THE LEUCOCYTE ANTIGEN *FactsBook*

A. Neil Barclay
Marian L. Birkeland
Marion H. Brown
Albertus D. Beyers
Simon J. Davis
Chamorro Somoza
Alan F. Williams

MRC Cellular Immunology Unit
Sir William Dunn School of Pathology
University of Oxford, Oxford, UK

Academic Press
Harcourt Brace Jovanovich, Publishers
LONDON SAN DIEGO NEW YORK BOSTON
SYDNEY TOKYO TORONTO

This book is printed on acid-free paper

ACADEMIC PRESS LIMITED
24-28 Oval Road
LONDON NW1 7DX

United States Edition published by
ACADEMIC PRESS INC.
San Diego, CA 92101

A catalogue record for this book is available from the British Li ary

ISBN 0-12-178180-1

Designed and typeset by Eric Drewery and Adrian Singer

Printed and bound in Great Britain by Mackays of Chatham c
Chatham, Kent

Contents

3 Protein Superfamilies and Cell Surface Molecules

CONCEPTS CONCERNING PROTEIN SUPERFAMILIES

Introduction

The amino acid sequences of most leucocyte surface proteins contain segments of sequence that have similarities to other proteins and it is likely that the similar sequences have been derived by divergent evolution from common precursors. Dayhoff et al.[1] introduced the terms "superfamily" for proteins with sequence similarity of 50% or less and "family" for those with more than 50% identity. In protein superfamilies there is often only 15–25% sequence identity and at this level it can be difficult to be confident that a sequence match indicates an evolutionary relationship, rather than just a chance similarity.

The first superfamily of leucocyte surface proteins to be defined was the immunoglobulin superfamily (IgSF) and this is now the largest with more than 100 different polypeptides on a variety of cell types [2]. The sequence identities between the members of this superfamily are at the 15–25% level but analysis showed that the conserved residues are clustered mainly in regions corresponding to the in-pointing residues of β strands of the Ig-fold. In contrast, the regions corresponding to the loops at the ends of the strands mostly show great sequence diversity. It may be regarded as a rule that in superfamilies of sequences that have derived by divergent evolution the conserved residues will relate to important structural features that are characteristic of the superfamily in question.

Several different superfamilies have been identified within leucocyte surface molecules. This chapter describes the methods for their identification and shows alignments of some sequences to illustrate the key residues that are often conserved in these superfamilies. A brief description of the structure and functions of each domain type is given.

Nomenclature for superfamilies, protein domains, repeats and motifs

There is no agreed nomenclature for most superfamilies and thus in this book we have tried to conform to the most commonly used names and in some cases to introduce abbreviations that might be useful. In general where superfamilies are named after a receptor we use the abbreviation "R". For example, the cytokine receptor superfamily is called the "cytokineR" superfamily. This seems useful in that the name becomes distinctive and distinguishes the superfamily usage from discussion in which a receptor is referred to in other ways. The naming of domains is a problem since one might discuss Ig domains either as domains of immunoglobulins or as domains of the superfamily. For the superfamily usage we include the abbreviation "SF" in cases where there may be ambiguities. For example, IgSF, scavengerRSF, FN type IIISF, CCPSF.

The term "domain" is used where it is likely that a segment of sequence forms a discrete structural unit, i.e. a peptide sequence whose three-dimensional conformation is not determined by other parts of the total protein sequence but is "self-contained". Three criteria are considered. First, proof of a domain structure comes from tertiary structure determination. Domains established at this level include: Ig, complement control protein (CCP), EGF, fibronectin (FN) type III, cytokineR and the C-type lectin. The MHC domain has also been revealed by X-ray

crystallography in MHC C
should not be referred to a
isolated unit rather than a
domains. However, we
segments as domains and
superfamily. The folds for
are discussed in the comm
Secondly, a domain str
that occur as the sole co
sequences) within protei
containing a high content
A third criterion for d
segments are found in th
other exons to form a ne
variety of structural dom
these last two criteria th
domains: FN type II, Link
In other cases it is n
structural unit, and for t
seen with the NGFR su
are always found togethe
pattern of exons does not
it appears that a precurs
repeat, and that addit
duplication and divergen
independent units withi
units interact to form a
"repeat" is used, include
The term "motif" is u
expected to form a folde
protein secretion and C
motifs, albeit of rather i
good example of a m
cytoplasmic domains of
signal transduction com
28.
The domains and repe
leucocyte surface molec
molecules and for surfa
kringle, thrombospondi

Identifying domains and
There can be problems
This is because the le
patterns are usually in s
the superfamily domai
program such as FASTA
a superfamily that the r
picks up all superfamil

crystallography in MHC Class I α1 and α2 domains and it might be argued that this should not be referred to as a domain since it is not clear that it will be found as an isolated unit rather than appearing always as a structural pair, as for the α1 and α2 domains. However, we will follow precedent in the field and refer to these segments as domains and to the proteins that show this fold as being in the MHC superfamily. The folds for all these domains are illustrated later in this chapter and are discussed in the commentary on each superfamily.

Secondly, a domain structure can also be argued for any superfamily segments that occur as the sole component of an extracellular sequence, or as sequence (or sequences) within proteins that is contiguous with hinge-like regions of sequence containing a high content of Ala, Gly, Pro, Ser and Thr residues.

A third criterion for defining a sequence as a domain is that the superfamily segments are found in the genome in single exons that can be readily spliced with other exons to form a new gene with an open reading frame. Proteins containing a variety of structural domains could then arise by recombination. On the basis of these last two criteria the following superfamilies can be referred to as containing domains: FN type II, Link, Ly-6, LDLR and the scavengerR.

In other cases it is not clear that a superfamily segment is an independent structural unit, and for these the term "repeat" is used. A good example of this is seen with the NGFR superfamily (Fig. 25), where a block of three or four repeats are always found together without intervening sequence between the repeats. The pattern of exons does not correlate with NGFR repeats. Thus with this superfamily it appears that a precursor structure evolved by gene duplication of a primordial repeat, and that additional members of the superfamily have evolved by duplication and divergence of the larger structure. The repeat segments may form independent units within the structure or alternatively it could be that the repeat units interact to form a larger structural unit. Superfamilies for which the term "repeat" is used, include NGFR and leucine-rich glycoprotein repeats.

The term "motif" is used to describe a smaller sequence pattern than might be expected to form a folded structural unit. Thus the patterns of signal sequences for protein secretion and GPI attachment (see Chapter 2) would be considered as motifs, albeit of rather ill-defined character in terms of sequence identities. A very good example of a motif is the conserved sequence pattern found in the cytoplasmic domains of the CD3, MB-1, and B29 antigens and other molecules of signal transduction complexes. Alignments identifying this motif are shown in Fig. 28.

The domains and repeats discussed in this chapter include only those present on leucocyte surface molecules. Additional domains have been described for secreted molecules and for surface molecules of other cell types. These include FN type I, kringle, thrombospondin, serine protease and perforin domains [3].

Identifying domains and repeats: testing the significance of relationships

There can be problems in identifying domains or repeats in new protein sequences. This is because the level of identities is often low and the conserved sequence patterns are usually in small patches throughout the 40–110 residues that make up the superfamily domains and repeats. A first step is to use a database searching program such as FASTA [4]. In many cases this will pick up some of the members of a superfamily that the new protein sequence matches. However, no search program picks up all superfamily members and it is not uncommon for a relationship to be

entirely missed. A second approach is to look by eye, or with a computer prog m [5], for the presence of sequence patterns that are characteristic of the dif rent superfamilies. These are thus noted in the sequence line-ups in this chapte For example, in the IgSF one would look for Cys residues with the patterns L/I/V- -L/I-X-C and D-X-G-X-Y-X-C for candidate regions that might occur aroun the conserved disulphide bond. In relation to these there should be other patch , for example V/L/Y-X-W corresponding to β-strand C. If the various conserved pa erns fall into place then a possible domain has been identified. The candidate d nain can be defined in relation to conserved sequence positions and then test l for statistical significance. For example, in the IgSF, the positions of the conserv . Cys residues, or equivalent residues if the domain lacks the typical disulphide bo l, are nominated and the domain is defined as beginning and ending 20 residues efore and after these positions. This proposed domain is then tested for the sta stical significance of sequence similarities against a set of domains that are accepte to be in the IgSF. For other superfamilies, other conserved residues would be chos n and the domain defined in relation to these. Possible key conserved residues are hown in the diagrams in this chapter and the designated residues are used to iden y the domains in the entries for the molecules.

In testing for statistical significance of a superfamily relationship it c ld be argued that the conserved pattern for a domain should be defined and the e ent to which this occurs in the new sequence should be tested. However, it an be difficult to define precisely a pattern for use in a statistical analysis since a many positions in the conserved pattern one of a group of alternative amino ac ls can occur. It is difficult to know how to treat sequence gaps in defining a patt n. For example, in the IgSF there can be very large differences in the length of the niddle of the domain and this creates problems in defining a pattern that is chara eristic of the IgSF to use in statistical analysis.

An alternative method to testing a sequence against a single sup family sequence pattern is to test it against a set of sequences (e.g. 20 sequences) hat are accepted as being members of the superfamily in question. In such an a lysis a simple statistical program that compares sequences pairwise for similarit can be used and the ALIGN program [1] has proved satisfactory for this purpose. n these comparisons no account is taken of superfamily patterns. However, if a se of good scores is obtained against a family of sequences, then the superfamily patt n must be present since this is the only pattern in common amongst the f nily of sequences against which the new domain is being tested.

In the ALIGN program of Dayhoff [1], the best alignment between two s uences is computed on the basis of a matrix of scores for all possible identities a . amino acid substitutions with a penalty scored each time a gap is introduced to mprove the number of good matches. The two sequences are then scrambled, reali ned and scored again to give a random best score and this is repeated 100–150 tin s. From the random scores a mean random score plus standard deviation (SD) is c culated and the test score is expressed in terms of its number of SDs above or l low the mean random score. Assuming a normal distribution and no effect due to election of sequences with particular compositions, values of 3.1, 4.3 and 5.5 D units indicate the probability of the sequence similarity arising by chance is 10^{-3}, 10^{-5} and 10^{-7} respectively. A score of 3 SD is considered to be a threshold for a ALIGN value that is of interest in indicating a superfamily relationship between p oteins.

The most common matrix of scores used with the ALIGN progra n is the

Table 1. *The 250 PAM Matrix of scores for sequence matches analysed by the ALIGN program adapted from ref. 107 with permission*

Table 1. *The 250 PAM Matrix of scores for sequence matches analysed by the ALIGN program adapted from ref. 107 with permission*

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W		
Cysteine (C)	12	0	-2	-3	-2	-3	-4	-5	-5	-5	-3	-4	-5	-5	-2	-6	-2	-4	0	-8	Cys	Sulphydryl
Serine (S)	0	2	1	1	1	1	1	0	0	-1	-1	0	0	-2	-1	-3	-1	-3	-3	-2	Ser	
Threonine (T)	-2	1	3	0	1	0	0	0	0	-1	-1	-1	0	-1	0	-2	0	-3	-3	-5	Thr	
Proline (P)	-3	1	0	6	1	-1	-1	-1	-1	0	0	0	-1	-2	-2	-3	-1	-5	-5	-6	Pro	Small hydrophilic
Alanine (A)	-2	1	1	1	2	1	0	0	0	0	-1	-2	-1	-1	-1	-2	0	-4	-3	-6	Ala	
Glycine (G)	-3	1	0	-1	1	5	0	1	0	-1	-2	-3	-2	-3	-3	-4	-1	-5	-5	-7	Gly	
Asparagine (N)	-4	1	0	-1	0	0	2	2	1	1	2	0	1	-2	-2	-3	-2	-4	-2	-4	Asn	
Aspartic (D)	-5	0	0	-1	0	1	2	4	3	2	1	-1	0	-3	-2	-4	-2	-6	-4	-7	Asp	Acid, acid amide
Glutamic (E)	-5	0	0	-1	0	0	1	3	4	2	1	-1	0	-2	-2	-3	-2	-5	-4	-7	Glu	hydrophilic
Glutamine (Q)	-5	-1	-1	0	0	-1	1	2	2	4	3	1	1	-1	-2	-2	-2	-5	-4	-5	Gln	
Histidine (H)	-3	-1	-1	0	-1	-2	2	1	1	3	6	2	0	-2	-2	-2	-2	-2	0	-3	His	
Arginine (R)	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6	3	0	-2	-3	-2	-4	-4	2	Arg	Basic
Lysine (K)	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5	0	-2	-3	-2	-5	-4	-3	Lys	
Methionine (M)	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6	2	4	2	0	-2	-4	Met	
Isoleucine (I)	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5	2	4	1	-1	-5	Ile	Small hydrophobic
Leucine (L)	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6	2	2	-1	-2	Leu	
Valine (V)	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4	-1	-2	-6	Val	
Phenylalanine (F)	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9	7	0	Phe	
Tyrosine (Y)	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	0	Tyr	Aromatic
Tryptophan (W)	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17	Trp	

The mutation matrix is based on the frequency of evolutionary replacements of one amino acid for another at homologous positions between present-day sequences and inferred ancestral sequences. One PAM unit is the unit of evolution represented by the matrix corresponding to one accepted amino acid substitution per 100 residues. This is discussed in detail in ref. 107.

Table 2. *ALIGN Scores for comparisons of IgSF domains*

V or V-related	β2 – microglobulin	Thy-1	N-CAM	CD48	CD45	C or C-related	β2 – microglobulin	Thy-1	N-CAM	CD48	CD45
Ig λ	-0.9	7.4	3.3	3.4	-0.6	Ig λ	5.6	1.4	4.7	0.5	-1.0
Ig κ	1.7	3.7	5.4	3.6	-0.1	Ig κ	6.0	1.3	4.0	0.4	-1.5
Ig heavy	1.1	3.9	3.9	4.0	1.5	Ig CH1	4.0	3.0	4.1	1.6	0.7
TcR β	1.8	3.3	4.6	3.6	-0.2	Ig CH2	2.4	2.9	3.8	0.0	0.0
TcR α	2.1	2.3	4.4	3.4	-1.5	Ig CH3	6.3	3.1	3.7	1.4	0.7
TcR γ	-0.2	1.6	3.9	3.3	1.4	TcR β	4.4	2.3	3.0	0.8	2.2
CD8 α	2.6	4.5	4.7	4.3	-0.3	TcR α	2.1	-0.3	1.7	0.3	-0.7
CD4 (1)	2.4	2.5	5.5	4.3	-0.1	TcR γ	1.9	0.8	3.6	1.0	-0.6
PolyIgR (1)	1.5	5.7	2.7	1.9	0.0	MHC I α3	8.2	2.2	2.9	2.4	1.3
PolyIgR (3)	1.7	5.8	4.3	2.4	0.9	MHC II α2	11.2	3.7	4.9	1.6	1.2
MRC OX-2 (1)	-1.1	5.0	5.3	3.3	-0.6	MHC II β2	11.3	2.4	4.3	1.1	1.6
Po protein	0.7	3.5	6.0	2.9	-0.4	CD1 α3	9.1	1.3	5.4	0.9	1.2

The IgSF domains were defined from a position 20 residues before the first Cys to 20 residues after the second Cys of a putative Ig-like disulphide bond or equivalent residues in the CD48 sequence. A sequence from rat CD45 is included as a control. This sequence shows none of the conserved patches of sequence characteristic of the IgSF domains but has two Cys and one Trp residue in approximately the same positions as in IgSF domains. The ALIGN program was run with the 250 PAMS Mutation Matrix, a bias of 6 and a break penalty of 6 and 150 random runs were performed. Details of the sequences are given in ref. 108.

mutation matrix calculated
replace each other in hom
substitution frequency expe
occurs for identities betwe
tryptophan and cysteine, th
observed to change more fre
mutation matrix are illustra
amino acids are favourabl
evolution. Those that repla
chance are shown in bold.

It is stated above that a
chance relationship of 10-5
comparisons with 20 me
superfamily relationship. H
and tests have been carried
program in analysis of the Ig
patterns were compared wit
roughly the correct position
IgSF patterns [6]. The results

Table 3. *ALIGN Scores for*

	GHR
GHR	1
PLR	**10.4**
GMP130	**4.0**
EPOR	**3.6**
IL3Rd1	**2.9**
GM-CSFR	**4.1**
IL6R	**4.5**
IL2Rβ	**2.2**
IL4R	**1.8**
IL3Rd3	**4.1**
IL7R	**1.3**

The regions of the domains sho
250 PAMS Mutation Matrix, a
scores in SD units are given.
putative cytokineRSF domain i

GHR, human growth hormo
GMP130, human membrane
receptor precursor; IL3Rd1 and
human GM-CSF receptor precu
receptor β chain precursor; IL4
precursor.

mutation matrix calculated from the observed frequencies at which am o acids replace each other in homologous proteins between species, compared ith the substitution frequency expected by chance. With this matrix a much hig :r score occurs for identities between residues that rarely change in evolution, such as tryptophan and cysteine, than for small residues like serine and threonin that are observed to change more frequently. The single letter code for amino acid and the mutation matrix are illustrated in Table 1. From this matrix it can be se ı which amino acids are favourable substitutes for one another in related pı :eins in evolution. Those that replace each other more often than would be ex :cted by chance are shown in bold.

It is stated above that an ALIGN score of 4.3 SD indicates a probal .ity of a chance relationship of 10^{-5} and thus it might be thought that one score of .3 SD in comparisons with 20 members of a superfamily would alone arş e for a superfamily relationship. However, theoretical probabilities may not l reliable and tests have been carried out to evaluate experimentally the use of tl ALIGN program in analysis of the IgSF. Sequences that were thought to have IgSl sequence patterns were compared with others that had two Cys residues and a Trp esidue in roughly the correct position for IgSF domains but otherwise did not sh v typical IgSF patterns [6]. The results are shown in Table 2 where it can be see that the

Table 3. *ALIGN Scores for alignments of cytokineRSF domains*

	GHR	PLR	GMP130	EPOR	IL3Rd1	GM-CSFR	IL6R	IL2Rβ	IL4R	IL3Rd3	IL7R
GHR		**10.4**	**4.0**	**3.6**	2.9	**4.1**	**4.5**	2.2	1.8	**4.1**	1.3
PLR	**10.4**		**8.1**	**7.2**	2.5	**7.3**	**5.2**	1.6	1.9	**4.4**	1.7
GMP130	**4.0**	**8.1**		**4.7**	2.4	**4.2**	**6.2**	**3.1**	2.5	**4.5**	1.1
EPOR	**3.6**	**7.2**	**4.7**		**4.7**	**5.6**	**5.1**	**3.6**	2.4	**8.7**	0.8
IL3Rd1	2.9	2.5	2.4	**4.7**		**4.3**	**3.1**	1.9	**3.0**	**5.3**	0.3
GM-CSFR	**4.1**	**7.3**	**4.2**	**5.6**	**4.3**		**5.3**	**5.5**	2.1	**6.7**	1.6
IL6R	**4.5**	**5.2**	**6.2**	**5.1**	**3.1**	**5.3**		**4.0**	1.7	**5.2**	1.1
IL2Rβ	2.2	1.6	**3.1**	**3.6**	1.9	**5.5**	**4.0**		2.5	**4.3**	0.6
IL4R	1.8	1.9	2.5	2.4	**3.0**	2.1	1.7	2.5		**3.8**	0.3
IL3Rd3	**4.1**	**4.4**	**4.5**	**8.7**	**5.3**	**6.7**	**5.2**	**4.3**	**3.8**		2.0
IL7R	1.3	1.7	1.1	0.8	0.3	1.6	1.1	0.6	0.3	2.0	

The regions of the domains shown in Fig. 4 were analysed using the ALIGN pro am with the 250 PAMS Mutation Matrix, a bias of 6, a gap penalty of 6 and 100 random ali ıments. The scores in SD units are given. Scores of 3 SD or greater are shown in bold. Note for the putative cytokineRSF domain in the IL7 receptor, all scores are less than 3 SD.

GHR, human growth hormone receptor precursor; PLR, rat prolactin recej ɔr precursor; GMP130, human membrane glycoprotein gp130 precursor; EPOR, mouse ythropoietin receptor precursor; IL3Rd1 and d3, mouse IL3 receptor precursor domains 1 an ; GM-CSFR, human GM-CSF receptor precursor; IL6R, human IL6 receptor precursor; IL21 , human IL2 receptor β chain precursor; IL4R, mouse IL4 receptor precursor; IL7R, huma IL7 receptor precursor.

control sequences give an occasional score above 2 SD but that no cc sistent pattern of good scores is obtained. In contrast the sequences that are consi ered to belong to the IgSF gave >40% scores of >3 SD. In practice, arguments foı ın IgSF relationship have proved reliable in terms of subsequent tertiary s ucture determination in cases where a group of scores >3 SD have been obtaiı d with >33% of one of the sets of IgSF sequences as defined below (i.e. the V set, (. set or C2 set). Convincing arguments are usually buttressed by one or two scores : >5 SD that are unlikely to arise by chance, even in isolation. A number of the .assical sequence patterns for the superfamily in question should be present in th correct positions in the sequence in relation to other conserved patches and the c ıserved sequences should be consistent with a structural prediction for the relevan lomain in cases where the domain structure is known. Thus there should be hyd phobic amino acids in the positions predicted to be in-pointing to stabilize the ter ıry fold and the Cys residues should potentially be able to form disulphide bonds hat are consistent with the fold. All these considerations were applied to the ar lysis of the IgSF relationship of the CD2 and CD4 antigens where there l s been controversy concerning CD2 domain 1 and CD4 domain 2. In both cas it was shown by structure determination that these domains were in the IgSF ıd that correct predictions were made for the β strands in both domains [7–10].

A further analysis using the ALIGN program is illustrated in Table : for the cytokineR superfamily. This is one of the most diverse superfamilies in erms of sequence alignments as can be seen from Fig. 3. The ALIGN scores clearl support the superfamily relationship for the grouped sequences with the excepti of the domains nominated for the IL4 and IL7 receptors. In the case of the IL4I domain only 2/10 scores are 3 SD or greater with another 4 scores being >2 SD. ıus the case for inclusion of the IL4R domain in the superfamily is weaker on th basis of the ALIGN scores. However, inspection of the conserved sequence patter in Fig. 3 leaves little doubt that the IL4R domain is in the cytokineR sup family. Cysteine residues can be confidently placed at all of the conserved posi ons and other conserved patterns are also present in the correct positions. For the roposed IL7R domain the situation is much more ambiguous. The ALIGN scores ıre very weak and only a hint of the conserved sequence patterns is seen in the ali ments. This domain would not be considered for inclusion in the cytokineR su rfamily except that the domain is present in a cytokine receptor. The case for thi domain will ultimately require validation by three-dimensional structure determir tion.

Domain sequence and structure: divergent and convergent evolution

In the above section criteria for defining a superfamily have been ısed on identifying a sequence pattern that is shared in a non-trivial way etween sequences of different molecules. It is then argued that the presenc of the sequence pattern indicates a relationship in evolution such that the dor ins that share the sequence pattern both derive from one original primordial lomain. However, it could be argued that a certain structure dictates a sequence pa ern and the sharing of the pattern is due to convergent evolution from different ecursor molecules rather than divergent evolution from a primordial domain. Con ersely it may be found that sequences with no detectable common pattern forı similar tertiary structures and thus that these are in the same superfamily eve though there is no detectable sequence relationship.

It now seems very unlikely that a general structure will dictate unique

sequence pattern. This can be seen from a consideration of sequences that can g e rise to domains with the Ig-fold. There are now five different sets of sequences w n. no convincing sequence similarity between them that can all give rise to the ;- fold. These are the sequences of the Ig superfamily, the FN type IIISF and · e cytokineRSF plus two sequences in the Pap-D bacterial protein that each give ı e to a domain with an Ig-fold [11-14]. There is also enormous diversity of seque .e within the IgSF that leads to the argument that there is no unique seque :e required to determine any part of the IgSF-fold. Thus it seems rather unlikely t ıt convergent evolution to yield the same structure would give rise to any comn n sequence pattern.

The converse argument is that all the sequences that give the same fold ıg structure have derived by divergent evolution and that all sequences with ı is structure should be included in the same superfamily. For example, the five set of sequences referred to above might all be considered as IgSF sequences. It does ot seem useful to take this point of view since there may be a relatively lim ed number of small stable protein folds that can occur and these may have evolvec on numerous occasions in evolution. In this case each of the sets of sequences v th the Ig-fold would have an independent primordial ancestor. Alternatively, tl re may have been one primordial structure which acquired mutations such that a ı w solution to the structure was produced, ultimately giving rise to sequences · at were not detectably similar to the ancestor family of sequences. At this stage tl re is no way to estimate the probability of the divergent versus the convergent · se for generation of the same structure without recognizable sequence similarity nd it seems best to stick to sequence patterns as the criteria for defining superfami es. This is sensible from a practical as well as a theoretical standpoint since sequ ce data are much more readily obtained than tertiary structural data and he superfamilies defined on the basis of sequence would be grouped as subsets wi in superfamilies based on tertiary structure considerations. It seems better to re in the sequence criterion and to note that certain superfamilies have the same fol ng patterns in their domains.

Given that the same structure can arise from various sequences, the ques on arises as to why sequence patterns are conserved in evolution. Molecules on he cell surface present unique determinants for interaction with a soluble mole le, the extracellular matrix or with other cell surface receptors. Such interact ons require diversity between molecules and not conservation of epitopes. he sequence patterns shared within a superfamily conserve the fold of the mol ule and usually involve residues pointing inwards in the folded structure rather ıan out-pointing residues that are available for biological interactions. Thus the question arises as to what evolutionary force can operate to preserve the ter ary structure of the molecule?

For cell surface molecules it can be argued that the key evolutionary pressu e is the requirement for molecular stability and, in particular, resistance to proteo sis. The small, tightly folded domains that make up most of the leucocyte mole ıles may have evolved as parts of stable coat proteins on single cell eukaryotes [6].] ese coat proteins then gave rise to the families of molecules that evolved along ith the evolution of multicellular organisms, to mediate cell division and regulati ı of cell differentiation. Surface molecules are generally resistant to prote ytic enzymes and this resistance is based on the folded structure, since dena ıred molecules are easily digested. One could argue that mutation to give ıew

recognition epitopes would be constrained by the necessity of preserving ıe folded structure of the domain. In general this led to preservation of certain sequence patterns that determine one particularly stable solution for the fold. umerous alternative sequence patterns may exist that could also give a stable f l, but to reach these a number of simultaneous mutations may be required an hence a switch to a new pattern may be a rare event in evolution. If a new patter did form this may become the founder of a new set of sequences in which the new attern is retained, again because of the pressure of proteolysis. From this viewpoi it seems likely that the Ig, FN type III and cytokineR superfamilies all arose from common ancestor via sequence shifts as described above. This view might b favoured because domains of these superfamilies are found in molecules wi similar functions and often a molecule may contain both Ig superfamily do ains and domains of the FN type III and cytokineR superfamilies. In particular, Ig and FN type IIISF domains are often found together in a single polypeptide.

Genomic structure and evolution of proteins with mixtures of domain ty s

The number of domains in a cell surface protein can vary greatly. In e Thy-1 antigen there is a single IgSF domain making up the whole of the ex acellular segment, whilst for the complement receptor 1 protein (CD35) the ex acellular region consists of 30 CCPSF domains in a linear array. In these cases nly one domain type is present but in other molecules there can be a mixture domain types. For example the L-selectin (LECAM-1) antigen contains C-type ectinSF, EGFSF and CCPSF domains.

The efficient build-up of proteins from individual domains during volution appears to depend on two aspects of genomic structure. There sho l be an approximate concordance of the domain ends with intron/exon boundari and the position of the intron with respect to the reading frame of a gene shoul be such that an open reading frame results from the recombination of an exo into the intron of an existing sequence [15]. Introns that are inserted after the first ase of a codon are called phase 1, those after the second base, phase 2, and thos between codons, phase 0 [16]. Analysis of the intron/exon boundaries of domains sent on leucocyte surface molecules shows that for most domain types each exon oundary is of the same phase (usually 1) as illustrated in Table 4. Recombinatio of such exons will lead to the construction of new open reading frames. The do in does not need to be contained within a single exon to allow shuffling as lo as the outermost intron boundaries are compatible. For instance, some IgSF do ains are coded for by two exons [2] and the cytokineRSF domain in the IL2 recepto s coded by three exons. In the latter case the internal splice sites are phase 2 an hase 0, whilst the external ones are phase 1; thus it is not possible to get p of the domain integrated into a sequence containing phase 1 splice sites.

THE SUPERFAMILIES THAT ARE FOUND IN LEUCOCYTE CE SURFACE MOLECULES

The superfamilies that are present in leucocyte surface molecules are scussed below together with alignments of some of the domains or repeat seque es. The alignments were made using a variety of computer programs (ALIGN [1], MPS [17], PILEUP [18]) and then modified after visual examination. The ends of the omains

Table 4. *Exo*

Domain or re

Complement
CytokineRSF
EGFSF
Fibronectin t
Fibronectin t
IgSF
Lectin C-type
Lectin C-type
Lectin S-type
Leucine-rich
LinkSF
LDLRSF
Ly-6SF
MHC
Nerve growt
ScavengerRS
Somatomedi

In both the Ig
two exons an
available on s
numbers of e
NK, not kno

can be diffi
consideratio
continues
highly like
domain 3 w
the domain
an asterisk
being a con
example, i
conserved
system the
structural
the sequen

Table 4. *Exon organization of domains and repeats of leucocyte surface mole les*

Domain or repeat type	Do the domain boundaries coincide with introns with same splice sites?	Splice site	Usual numbe of exons r domai
Complement control protein (CCP)SF	Yes	type 1	1
CytokineRSF	Yes	type 1	2
EGFSF	Yes	type 1	1
Fibronectin type IISF	Yes	type 1	1
Fibronectin type IIISF	Yes	type 1	1
IgSF	Yes	type 1	1 or 2
Lectin C-typeSF (e.g. selectins)	Yes	type 1	1
Lectin C-typeSF (e.g. Kupffer cell receptor)	No	NA	3
Lectin S-typeSF	NK	NK	NK
Leucine-rich glycoprotein repeat	No	NA	NA
LinkSF	Yes	type 1	1
LDLRSF	Yes	type 1	1
Ly-6SF	No	NA	NA
MHC	Yes	type 1	1
Nerve growth factor receptor (NGFR)SF	No	NA	NA
ScavengerRSF	NK	NK	NK
Somatomedin BSF	Yes	type 1	1

In both the IgSF and the CCPSF domains there are examples where the domain is enco d by two exons and also where two domains are encoded by one exon. Only limited data are available on some of the domains and it is possible that other examples with different numbers of exons per domain or motif may be found.
NK, not known, NA, not applicable.

can be difficult to define from the sequence and this problem is illustra d by consideration of the structure for CD4. In CD4 the last β strand of don in 1 continues directly into domain 2 and between CD4 domains 3 and 4 it ems highly likely that the overlap will be even greater and that the last β str d of domain 3 will also be the first β strand of domain 4. Thus in the alignments s own, the domains are defined with respect to key internal residues that are marke with an asterisk, and the beginnings and ends can be taken for statistical comparis ns as being a constant number of residues before and after the conserved positio . For example, in the case of the IgSF this is taken as 20 residues before and af r the conserved Cys positions. If the goal was to express a single domain in an exp ssion system then sequence alignments and structure should be taken into accoun nd a structural prediction would be attempted on the basis of all the data to dec le on the sequence that should be expressed.

The complement control protein (CCP) superfamily (Figs 1 and 2)
This domain is named CCP because it is commonly found in proteins that control the complement cascade [19]. For instance, factor H consists solely of 20 CCPSF domains whilst other complement components contain CCPSF domains mixed with other domains, e.g factors B and C2 each contain three CCPSF domains together with a serine protease domain. The CCP domain is also commonly called the short consensus repeat or SCR [19]. It is present in widely different numbers in cell surface molecules ranging from 30 domains in complement receptor 1 (CD35) to a single domain in L-selectin. These domains are clearly involved in protein binding and the CR1 (CD35) and CR2 (CD21) complement binding regions have been mapped to the first four CCP domains of each of the first three groups of seven domains in CD35 and to the first two domains of CD21.

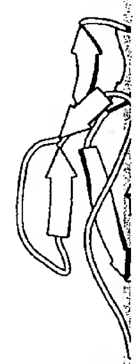
The structure of one CCPSF domain from complement control protein factor H has recently been solved using NMR and consists of two segments of antiparallel β sheet and a short triple-stranded β sheet with no α-helical structure [20]. The folding pattern for this domain is shown in Fig. 2 and the b strand positions are marked above the sequence alignments shown in Fig. 1 .

Cytokine receptor (cytokineR) superfamily (Figs 3 and 4)
Three domain types are found in cytokine receptors including those of the Ig, FN type III and cytokineR superfamilies. A common arrangement is to have a single NH_2-terminal cytokineRSF domain followed by an FN type IIISF domain, but there are variations on this theme. Initially these two domain types were not distinguished [5] and the term haematopoietin receptor superfamily was widely used for molecules containing this pair of domain types [21,22]. We use the term cytokine receptor superfamily for the domain of about 100 amino acids usually found NH_2-terminal to the FN type IIISF domain and alignments of domains from this superfamily are shown in Fig. 3. Analysis with the ALIGN program (Table 3) gives good evidence for the presence of the cytokineRSF domain in the receptors for IL2 (β chain), IL3, IL6, growth hormone, granulocyte-macrophage colony stimulating factor, erythropoietin and in the GMP130 protein [22]. The presence of a cytokineRSF domain in IL4R is less strongly supported by ALIGN analysis but as discussed above the case for inclusion of this domain is convincing if all the data are considered. The IL7 receptor contains a clear FN type IIISF domain but the sequence at the NH_2-terminal region shows only a distant relationship to the cytokineRSF domains (see p.338). The possible cytokineR domain in the IL7 receptor [22,23] is shown below the other sequences in Fig. 3 but the correctness or otherwise of this assignment will require validation by tertiary structure determination.

The structure of the growth hormone receptor has recently been solved by X-ray crystallography [13] and this has revealed the fold for the cytokineRSF and the FN type IIISF domains that constitute the extracellular domain of this receptor (an FN type IIISF domain has also been solved by NMR – see below). These domains have similar folds that are also similar to the folds of IgSF C2 set domains [8,9] and the PapD chaperone protein domains [11]. Bazan [24] had previously argued that there may be structural similarities between cytokineRSF domains, FN type IIISF domains and IgSF domains on the basis of predicting patterns of β strands in the sequences. Despite the success of these predictions the degree of sequence similarity between these domain types is low. The cytokineRSF domains have a characteristic Cys-X-Trp sequence together with three other conserved Cys residues, whilst the FN type

Factor H	L P C K S - P P E
CD35d12	R V C Q P - P P D
Factor B	G S C S L - - - E
L-Selectin	I Q C E - - P L E
C4BPA	N S C I N - L P D
IL2R1	E L C D D D P P E
FXIII	E P C T V - N V D
	*
Factor H	F G I D G P A -
CD35d12	Y D L R G A A -
Factor B	F Y P Y P V Q -
L-Selectin	T N L T G I E -
C4BPA	Y K P T T D E P
IL2R1	F R R I K S G S L
FXIII	Y D L S P L T P L

Figure 1. *CCP superfamily d... are boxed. The lines above t... strands determined from th... (see Fig. 2) [20]. The asterisks... shown on the Figures to ide... the entries in Section II. Th... Swissprot database unless o... number and residue numbe... factor H precursor domain ... 1 precursor domain 12 (P17... factor B (P00751, 10–74); L-... C4BPA, complement C4-bi... receptor α chain precursor (... precursor (P05160, 452–516)*

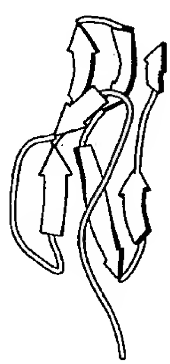


Factor H CCPSF

```
Factor H     L P C K S - P P E I S H G V V A H M - - - - - S D S Y Q Y G E E V T Y   F E G
CD35d12      R V C Q P - P P D V L H A E R T Q R D K - - - - D N F S P G Q E V F Y   E P G
Factor B     G S C S L - - - E G V E I K G G S F R L - - - - - - L Q E G Q A L E Y   P S G
L-Selectin   I Q C E - - P L E A P E L G T M D C T H P - - L G N F N F N S Q C A F   S E G
C4BPA        N S C I N - L P D I P H A S W E T Y P R P T K E D V Y V V G T V L R Y   H P G
IL2R1        E L C D D D P P E I P H A T F K A M - - - - - - - A Y K E G T M L N C I   K R G
FXIII        E P C T V - N V D Y M N R N N I E M K W - K Y E G K V L H G D L I D F   K Q G
                 *

Factor H     F G I D G P A - - - - - I A K C L G - E K W S H P - - - - - - - - - - -   C I
CD35d12      Y D L R G A A - - - - - S M R C T P Q G D W S P A A - - - - - - - - -   C E
Factor B     F Y P Y P V Q - - - - - T R T C R S T G S W S T L K T Q D Q K T V R K   C R
L-Selectin   T N L T G I E - - - - - E T T C E P F G N W S S P E - - - - - - - - -   C Q
C4BPA        Y K P T T D E P T - - - T V I C Q K N L R W T P Y Q - - - - - - - - - -   C E
IL2R1        F R R I K S G S L - - - Y M L C T G N S S H S S W D N Q C - -   - - -   C T
FXIII        Y D L S P L T P L S E L S V Q C N R - G E V K Y - - - - - - - - - - -   C T
                                                                                   *
```

Figure 1. *CCP superfamily domains. Residues identical in four or more se uences are boxed. The lines above the sequences correspond to the positions of tl β strands determined from the structure of factor H domain 16, residues 92 -985 (see Fig. 2)* [20]*. The asterisks mark the positions of the conserved residues at are shown on the Figures to identify domains in each entry for a molecule as own in the entries in Section II. The sequences of the following proteins are from he Swissprot database unless otherwise indicated and the database accessio number and residue numbers are given in brackets. Factor H, human con lement factor H precursor domain 16 (P08603, 929–985); CD35d12, complement ceptor 1 precursor domain 12 (P17927, 745–799); Factor B, HR16 human comple ent factor B (P00751, 10–74); L-selectin, L-selectin precursor (P14151, 195–25 C4BPA, complement C4-binding protein (P04003, 249–313); IL2R1, interl kin-2 receptor α chain precursor (P01589, 22–83); FXIII, coagulation factor XIII chain precursor (P05160, 452–516).*



Factor H CCPSF domain

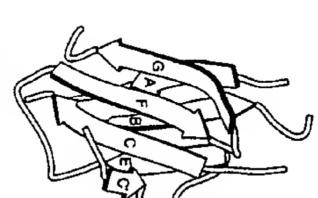
Figure 2. *The folding pattern of CCPSF domain. Ribbon diagram showi the folding pattern of a CCPSF dom n from factor H determined by NMR* [20]*. he β strands are shown as broad arro s pointing from the amino to carb y direction and the connecting lo s as thinner lines.*

Figure 3. Cytokine receptor superfamily

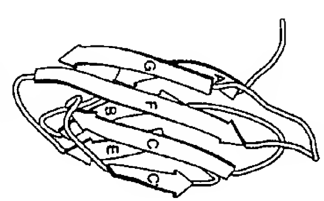
```
                        A                         B                               C                      C'
GHR     NSSKEPKFTKCRSPE-----RETFSCHWTDEVHHGTKNLGPIQLFYTRRNTQEWTQEWKECPDYV
PLR     SPPGKPEIHKCRSPD-----KETFTCWWNPGTDGGLP--TNYSLTYSKEGEKT----TYECPDYK
GMP130  LPPEKPKNLSCIVNE-----GKKMRCEWDGGRETHL--ETNFTLKSEWATH-----KFADCKAK-
EPOR    LASRGSEELLCFTQR-----LEDLVCFWEEAASSGM--DFNYSFSYQLEG-----ESRKSCSLHQ
IL3Rd3  GDKAQPQNLQCFFDG-----IQSLHCSWEVWTQTTG--SVSFGLFYRPSPA----APEEKCSPVV
GMCSFR  REGTAAQNFSCFIYN-----ADLMNCTWARGPTAP--RDVQYFLYIRNSK----RRREIRCPYYI
IL6R    VPPEEPQ-LSCFRKSP----LSNVVCEWGPRSTPSL--TTKAVLLVRKFQNSPAEDFQEPCQYSQ
IL2Rb   AAVNGTSQFTCFYNS-----RANISCVWSQDGALQD---TSCQVHAWPDR----RRWNQTCELLP
IL4R    GSIKVLGEPTCFSDY-----IRTSTCEWFLDSAVDC--SSQLCLHYRLMFFE--FSENLTCIPRN
IL3Rd1  EETVPLKTLQCYNDY-----TNRIICSWADTEDAQG--LINMTLLYHQLDK----IQSVSCELSE
                                  *
IL7R    DAELDDYSFSCYSQLEVNGSQHSLTCAFEDPD-----VNTTNLEFEICGA-----LVEVKCLNFR

                     E                F                G
GHR     -------SAGENSCYF--NSSFTSIWIPYCIKL--TSNGGTVDEKCFSVDEIV
PLR     -------TSGPNSCFF--SKQYTSIWKIYIITVNATNQMGSSSSDPLYVDVTY
GMP130  -------RDTPTSCTV---DYSTVYFVNIEVWVEAENALGKVTSDHINFDPVY
EPOR    -APTVRGSVRFWCSL--PTADTSSFVPLELQV--TEASGSPRYHRIIHINEV
IL3Rd3  KEPQASVYTRYRCSL--PVPEPSAHSQYTVSVK---HLEQGKFIMSYYHIQM
GMCSFR  -----QDSGTHVGCHL--DNLSGLTSRNYFLVNGTSREIGIQFFDSLLDTKKI
IL6R    -------ESQKFSCQL---AVPEGDSSFYIVSMCVASSVGSKFSKTQTFQGCG
IL2Rb   ------VSQASWACNLILGAPDSQKLTTVDIVTLRVLCREGVRWRVMAIQDFK
IL4R    ------SASTVCVCHM--EMNRPVQSDRYQMEL--WAEHRQLWQGSFSPSGNV
IL3Rd1  -------KLMWSECPSSHRCVPRRCVIPYTRFSNGDNDYYSFQPDRDLGIQLM
                     *
IL7R    -------KLQEIYFIET-KKFLLIGKSNICVKVGEKSLTCK-KID---LTTIV
```


Figure 3. (opposite) *Cytokin*
more sequences are boxed.
residues that are shown on
molecule as shown in Sectio
the Swissprot database unle
number and residue numbe
receptor precursor (P10912,
21–116); GMP130, human n
124–218); EPOR, mouse ery
mouse IL3 receptor precurs
243–347); GMCSFR, human
IL6 receptor precursor (P08
precursor (P14784, 26–125);
IL7R, human IL7 receptor p
are from 20 amino acids N
start corresponds to residue
more difficult to define due
close to the predicted boun
type IIISF domains in GHR



Human GHR domain 1



Fibronectin domain 21

Figure 3. (opposite) *CytokineR superfamily domains. Residues identical in four o more sequences are boxed. The asterisks mark the positions of the conserved residues that are shown on the figures to identify domains in each entry for a molecule as shown in Section II. The sequences of the following proteins are froı the Swissprot database unless otherwise indicated and the database accession number and residue numbers are given in brackets. GHR, human growth hormo receptor precursor (P10912, 46–147); PLR, rat prolactin receptor precursor (P0571 21–116); GMP130, human membrane glycoprotein gp130 precursor (PIR: A36337 124–218); EPOR, mouse erythropoietin receptor precursor (P14753, 42–140); IL3I mouse IL3 receptor precursor domains 1 and 3 (PIR: A35782, d1, 29–127; d3 243–347); GMCSFR, human GM-CSFR precursor (P15509, 116–214); IL6R, huma IL6 receptor precursor (P08887, 112–214); IL2Rb, human IL2 receptor β chain precursor (P14784, 26–125); IL4R, mouse IL4 receptor precursor (P16382, 24–122, IL7R, human IL7 receptor precursor (P16871, 32–127). The sequence alignments are from 20 amino acids NH_2-terminal from the conserved CXW. The sequence start corresponds to residue 2 in the prolactin receptor. The COOH-terminus is more difficult to define due to the lack of conserved residues and that shown is close to the predicted boundary between the cytokineRSF domains and the FN type IIISF domains in GHR, PLR, IL6R.*

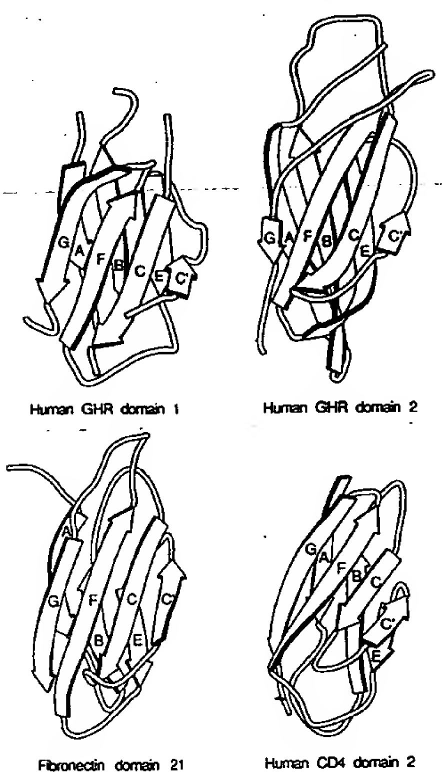


Figure 4. *The folding pattern of the cytokineR superfamily and fibronectin type IIISF domains. Ribbon diagrams f the cytokineR superfamily a FN type IIISF domains from human growth hormone receptor [13], and FN type IIIS domain 21 from human fibronectin [14]. The IgSF C2-s domain from human CD4 domain 2 is included for comparison [8,9]. The β strand are shown as broad arrows pointing from the amino to carboxy direction and the connecting loops as thinner lines. Some gaps are present the loops of the growth hormone receptor where the structure has not been fully resolved [13]. Each β strand is labelled using the same nomenclature as in the IgSF This lettering corresponds to that in the sequence alignments (Figs 3,8,12).*

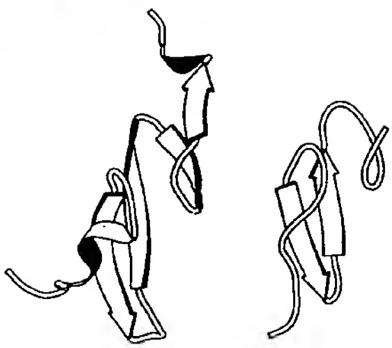
IIISF domains lack a conserved pattern of Cys residues. Using the ALIGN program 100 comparisons between cytokineRSF and FN type IIISF domains were made and of these 81 gave a score less than 2 SD and only eight comparisons gave scores greater than 3.0 SD (5.9, 5.6, 4.2, 3.7, 3.5, 3.4, 3.3, 3.0 SD; unpublished observations). Although there were two good scores out of 100 and some moderate ones, the case for a relationship in evolution based on ALIGN analysis is much weaker than that for members of the individual superfamilies. The possibility of an origin by divergent evolution for the cytokineR, FN type IIISF and IgSF domains is discussed above.

Several members of the cytokineR superfamily show small patches of sequence similarity in their cytoplasmic domains. These are reviewed in ref. 25.

Epidermal growth factor (EGF) superfamily (Figs 5 and 6)
EGFSF domains are found in EGF itself and in transforming growth factor (TGF) α. This domain is also found in a variety of secreted proteins such as blood coagulation factor IX and cell surface molecules such as in the selectins L-selectin, E-selectin and P-selectin (CD62). The structures of EGF, TGFα and the factor IX EGFSF domain have recently been determined and show similarity in folding pattern [26–28]. The structures of EGF and factor IX EGFSF domains are shown in Fig. 6. The latter is slightly smaller than EGF itself but is probably representative of the repeating EGFSF domains found in many proteins (see Fig. 5). The single EGFSF domain from factor IX has functional activities distinct from the EGF itself, for example it has Ca^{2+}-binding activity [28]. It is likely that EGFSF domains are

```
FA9-1   V D G D Q C - - - E S N P C L N G G S C K D - - D I N S Y E C W C P F G F E G K - - - - N C E L
FA9-2   - - D V T C N I - K N G R C - - E Q F C K N S - A D N K V V C S C T E G Y R L A E N Q K S C E P
EGF     N S D S E C P L S H D G Y C L H D G V C M Y I E A L D K Y A C N C V V G Y I G E - - - - R C Q Y
L-Sel   - - T A S C - - - Q P W S C S G H G E C V E - - I I N N Y T C N C D V G Y Y G P - - - - Q C Q F
CD62    - - T A S C - - - Q D M S C S K Q G E C L E - - T I G N Y T C S C Y P G F Y G P - - - - E C E Y
E-Sel   - - T A A C - - - T N T S C S G H G E C V E - - T I N N Y T C K C D P G F S G L - - - - K C E Q
PRTC    P L E H P C - - - - A S L C C G H G T C I D - - G I G S F S C D C R S G W E G R - - - - F C Q R
114/A10 G P S D L C - - - N P N P C K G T A S C V K - - L H S K H F C L C L E G Y Y Y N S S L S S C V K
NOTCH   T N D E D C - - - T E S S C L N G G S C I D - - G I N G Y N C S C L A G Y S G A - - - - N C Q Y
                  *                                                                             *
```

Figure 5. *EGF superfamily domains. Residues identical in five or more sequences are boxed. The asterisks mark the positions of the conserved residues that are marked on the domain organization figures in the entries in Section II. The sequences of the following proteins are from the Swissprot database and the database accession number and residue numbers are given in brackets. FA9-1 and FA9-2, human coagulation factor IX precursor (P00740, 92–130 and 131–172); EGF, human epidermal growth factor precursor (P01133, 971–1014); L-Sel, human L-selectin precursor (P14151, 157–193); CD62, human CD62 or P-selectin precursor (P16109, 160–196); E-Sel, E-selectin precursor (P16581, 138–176); PRTC, human protein C precursor (P04070, 96–133); 114/A10, mouse haematopoietic cell surface protein 114/A10 precursor (P19467, 232–274); NOTCH, Drosophila notch protein (P07207, 1021–1059). The ends of the alignment correspond to those of the coagulation factor IX EGFSF domain whose structure has been determined [28]. The structure of EGF itself has been determined for a sequence that extends a further four residues beyond that shown (see Fig. 6) [27]. The bars above the sequence indicate the positions of the β-strands.*



EGF

Factor XI EGFSF domain

Figure 6. *The folding pattern of EGFSF domains. Ribbon diagrams of EGF[27] and a coagulation factor IX EGFSF domain[28]. The β strands are shown as broad arrows pointing from the amino to carboxy direction and the connecting loops as thinner lines. The NH_2-terminal core of the structure is similar in both domains but the EGF structure extends further with two more short β strands.*

```
Fibr d7     T A V T Q T Y G G N S N G E P C V L P F T Y N G R T F Y S
Fibr d8     T V L V Q T Q G G N S N G A L C H F P F L Y N N H N Y T D
Mannose R   Y E A M Y T L L G N A N G A T C A F P F K F E N K W Y A D
Factor XII  K A E E H T V V L T V T G E P C H F P F Q Y H R Q L Y H K
Collag      R A D S T V M G G N S A G E L C V F P F T F L G K E Y S T

                                                    *
C T T E G R Q D G H L W C S T T S N Y E Q D Q K Y S F C T D H T
C T S E G R R D N M K W C G T T Q N Y D A D Q K F G F C P M A A
C T S A G R S D G W L W C G T T T D Y D T D K L F G Y C P L K F
C T H K G R P G P Q P W C A T T P N F D Q D Q R W G Y C L E P K
C T S E G R G D G R L W C A T T S N F D S D K K W G F C P D Q G
```

Figure 7. *Fibronectin type IISF domains. Residues identical in three or more sequences are boxed. The asterisks mark the positions of the conserved residues that are marked on the domain organization figures in the entries in Section II. The sequences of the following proteins are from the Swissprot database and the database accession number and residue numbers are given in brackets. Fibr, human fibronectin precursor (P02751; d7, 314–373; d8, 374–434); Mannose R, human mannose receptor precursor (P22897, 153–212); Factor XII, human coagulation factor XII precursor (P00748, 32–91); Collag, human type V collagenase precursor (EC 3.4.24.7) (P14780, 332–391). The ends of the alignment are based on the exon boundaries of the fibronectin domains.*

recognition structures that can be involved in various functions. For example, tv of the 36 EGFSF domains of the *Drosophila* protein *Notch* have been shown to necessary for the interaction of *Notch* with the *Delta* and *Serrate* proteins[29].

Fibronectin (FN) type II superfamily (Fig. 7)

The FN type II domains were first identified as one of three different repeati sequence patterns within the fibronectin molecule. The FN type IISF domain h

been found in few other proteins and the only leucocyte molecule with this do ain is the mannose receptor which contains one FN type IISF domain. The structu of a sequence from bovine seminal fluid protein PDC-109 that shows sequ ce similarity over part of the FN type II domain alignment shown in Fig. 7, has l en determined by NMR [30].

Fibronectin (FN) type III superfamily (Figs 4 and 8)

The FN type IIISF domain was identified in an extracellular matrix protein b is also common in membrane molecules and particularly these found in the ner us system which often have IgSF domains [2]. It has also been found in large numbe in the group of muscle proteins that bind myosin such as twitchin in *Caenorhab tis elegans* [31] and also in titin in mammals [32]. This is the only group of IgSF molec es found so far in the cytosol. Another example of cytosolic localization of FN pe IIISF domains is in the cytoplasmic segment of the integrin β_4 chain [33] (note he external regions of integrins do not contain any FN type III SF domains). Th is currently the only example of a domain found at the surface of leucocytes whic is also present on the cytoplasmic side of a transmembrane protein.

Structures for FN type IIISF domains have recently been solved by NMR [14] nd X-ray crystallography [13]. This domain consists of two β sheets with a sin ar folding pattern to the IgSF fold, the CytokineR domain and the domains of he PapD chaperone protein [11]. However, there is no significant sequence simil ty amongst these proteins as analysed by the methods discussed above.

Immunoglobulin (Ig) superfamily (Figs 9–12)

The immunoglobulin superfamily (IgSF) is the largest superfamily of cell sur ce proteins in general and for leucocyte antigens in particular, as is evident from e collated data in Table 1 in Chapter 1 which shows that approximately 40% of leucocyte membrane polypeptides contain IgSF domains. The structures of sev al IgSF domains have been determined by X-ray crystallography including Ig V- an C-

Figure 8. (opposite) *Fibronectin type IIISF domains. Residues identical in five or more sequences are boxed. The positions of the β strands determined for domai 21 of human fibronectin are indicated above the sequences [14]. See Fig. 4 for folding patterns of FN type IIISF domains from fibronectin and growth hormon receptor. The asterisks mark the positions of the conserved residues that are marked on the domain organization figures in the entries in Section II. The sequences of the following proteins are from the Swissprot database unless otherwise indicated and the database accession number and residue numbers a given in brackets. GHR, human growth hormone receptor precursor (P10912, 148–251); FIBR, human fibronectin precursor (P02751:d12 605–700, d13 719–80 d16 996–1085, d21 1447–1541); LAR, human LAR precursor (P10586, 596–694); TWIT, twitchin cytoplasmic protein from* Caenorhabditis elegans *(PIR:S07571 1761–1854); CAML1, mouse neural adhesion molecule L1 precursor (P11627, 916–1012); IL7R, human interleukin 7 receptor precursor (P16871, 128–231); GMP130, human membrane glycoprotein gp130 precursor (PIR: A36337, 221–32 PLR, rat prolactin receptor precursor (P05710, 121–224); IL3LR, mouse IL3-receptor-like protein precursor (AIC2B) (PIR: A35782, d2; 135–243, d4; 342–441)*

Figure 8. Fibronectin type IIISF domains

```
                   A                 B                              C
FIBR-21   VSDVPRDLEVVAATPT-------SLLISWDAPAVT----------VRYYRITYGETG
GHR       QPDPPIALNWTLLNVSLTGIHADIQVRWEAPRNADIQK--GWMVLEYELQYKEVN
FIBR-12   YPSSSGPVEVFITETPS--QPNSHPIQWNAPQPSH---------ISKYILRWRPKN
FIBR-13   SPLVATSESVTEITA-------SSFVVSWVSASDT----------VSGFRVEYELS-
FIBR-16   KLDAPTNLQFVNET-------DSTVLVRWTPPRAQ----------ITGYRLTVGLTR
LAR       PSAPPQKVMCVSMGS-------TTVRVSWVPPPADSRNG----VITQYSVAHEAVD
TWIT      RPDRPGRPEPTDWDS-------DHVDLKWDPPLSDGGA-----PIEEYQIEKRTKY
CAML1     VPGHPEALHLECQSD-------TSLLLHWQPPLSHNG------VLTGYLLSYHPVE
IL7R      KPEAPFDLSVIYREGA------NDFVVTFNTSHLQKKYV----KVLMHDVAYRQEK
GMP130    KPNPPHNLSVINSEELS-----SILKLTWTNPSIKSV------IILKYNIQYRTKD
PLR       IVEPPRNLTLEVKQLKD--KKTYLWVKWSPPTITDVKT--GWFTMEYEIRLKPEE
IL3LRd2   QPPLPKNVSISSSED---------RFLLEWSVSLGDAQVSWLSSKDIEFEVAYKRLQ
IL3LRd4   IQMEPPTLNLTKNRD---------SYSLHWETQKMAY-----SFIEHTFQVQYKKKS
                                     *

              C'            E             F                         G
FIBR-21   GNSPVQEFTVPGS---KSTATISGLKPGVDYTITVYAVTGRGDSP-----ASSKPISINYRTE
GHR       -ETKWKMMDPIL-----TTSVPVYSLKVDKEYEVRVRSKQRNSG--------NYGEFSEVLYVTL
FIBR-12   SVGRWKEATIPG---HLNSYTIKGLKPGVVYEGQLISIQQYGH----------QEVTRFDFTTT
FIBR-13   EEGDEPQYLDLPS--TATSVNIPDLLPGRKYIVNVYQISEDGE----------QSLILSTSQTT
FIBR-16   -RGQPRQYNVGPS---VSKYPLRNLQPASEYTVSLVAIKGNQE----------SPKATGVFTTL
LAR       GEDRGRHVVDGIS-REHSSWDLVGLEKWTEYRVWVRAHTDVGPG---------PESSPVLVRTD
TWIT      --GRWEPITVPG---GQTTATVPDLTPNEEYEFRVVAVNKGGP----------SDPSDASKAVI
CAML1     GESKEQLFFNLSDP-ELRTHNLTNLNPDLQYRFQLQATTQQGG--    ----PGEAIVREGGTM
IL7R      DENKWTHVNLSST---KLTLLQRKLQPAAMYEIKVRSIPDHYF--KGFWSEWSPSYYFRTPEI
GMP130    -ASTWSQIPPEDTASTRSSFTVQDLKPFTEYVFRIRCMKEDG---KGYWSDWSEEASGITYED
PLR       -AEEWEIHFTG----HQTQFKVFDLYPGQKYLVQTRCKPDHG-----YWSRWSQESSVEMPND
IL3LRd2   --DSWEDAYSLHTSKFQVNFEPKLFLPNSIYAPRVRTRLYPGSSLSGRPSRWSPEAHWDSQPQ
IL3LRd4   --DSWEDSKTENL-DRAHSMDLSQLEPDTSYCARVRVKPISNY--DGIWSKWSEEYTWKTDWV
                                         *
```

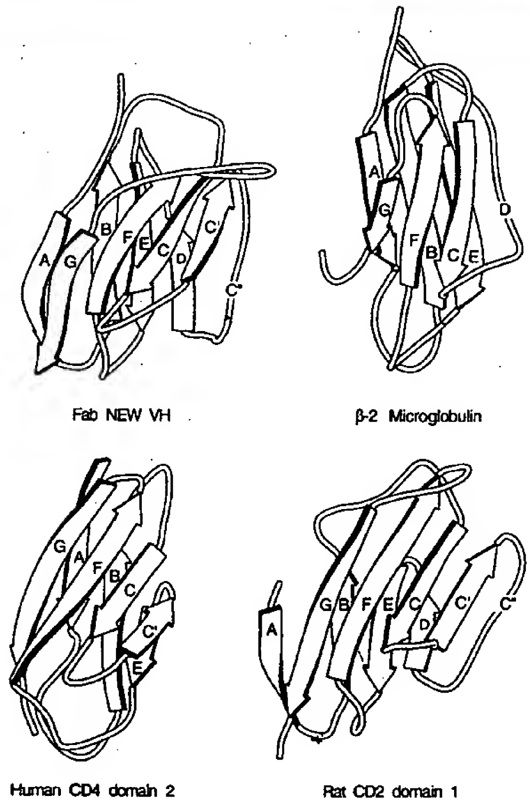


Figure 9. *The folding pattern of IgSF domains. Ribbon diagrams for IgSF domains. Ig V set (V_H of human NEW Fab); Ig C1 set (β 2-microglobulin); Ig C2 set (CD4 domain 2); and Ig V set lacking the normally conserved disulphide between β strands B and F (rat CD2 domain 1). The β strands are shown as broad arrows pointing from the amino to carboxy direction and the connecting loops as thinner lines. These are labelled with the corresponding strand letters used in the alignments for the Ig V-set, C1-set and C2-set sequences (Figs 10–12) and in the FN typeIIISF domains (Figs 4 and 8). The data are from the Brookhaven Protein Structure Database apart from CD2* [10]*.*

Figure 10. (opposite) *Immunoglobulin V-set domains. Residues identical in five or more sequences are boxed. The positions of the β strands are indicated above the sequences. The asterisks mark the positions of the conserved residues that are marked on the domain organization figures in the entries in Section II. The sequences of the following proteins are from the Swissprot database and the database accession number and residue numbers are given in brackets. Ig lambda, mouse Ig λ chain precursor (MOPC 104E) (P01724, 21–129); Ig kappa, human Ig κ chain Roy (P01608, 3–107); Ig heavy, human Ig heavy chain NEWM (P01825, 3–116); TcR beta, human TcR β chain precursor (P01733, 22–135); TcR alpha, mouse TcR α chain precursor (P01739, 23–132); CD8 beta, rat CD8 β chain precursor (P05541, 21–134); CD8 alpha, rat CD8 α chain precursor (P07725, 27–138); CD4d1, human CD4 precursor domain 1 (P01730, 21–123); Thy-1, rat Thy-1 precursor (P01830, 18–128); CD2 d1, rat CD2 precursor domain 1 (P08921, 20–120).*

Figure 10. Immunoglobulin V-set domains

```
                 A              B                 C              C'           C"
Ig lambda   AVVTQESALTTSPGETVTLTCRSSTGAV TTSNYANWVQQKP--DHLFTGLIGG--TNNRAPGV---
Ig kappa    QMTQSPSSLSASVGDRVTITCQASQD----ISIFLNWYQQKPG-KAPKLLIYDA---SKLEAGV---
IgG heavy   QLEQSGPGLVRPSQT-LSLTCTVSGS--TFSNDYYTWVRQPPG-RGLEWIGYVFYH--GTSDDTTPL
TcR beta    GVIQSPRHEVTEMGQEVTLRCKPISGH-----NSLFWYRQTMM-RGLELLIYFN--NNVPIDDSGMP
TcR alpha   NVQQSPESLIVPEGARTSLNCTFSDS----ASQYFWWYRQHSG-KAPKALMSIFS--NGEKE-----
CD8 beta    ALLQTPSSLLVQTNQTAKMSCEAKTFPK---GTTIYWLRELQDSNKNKHFEFLASRTSTKGIKY---
CD8 alpha   QLQLSPKKVDAEIGQEVKLTCEVLRDTS----QGCSWLFRNSSSELLQPTFIIYVSSSRSKLNDILD
CD4 d1      AATQGKKVVLGKKGDTVELTCTASQKK----SIQPFHWKNSN-----QIKILGNQG---SFLTKGPSKL
Thy-1       RGQRVISLTACLVNQNLRLDCRHENNTNLPIQHEFSLTRE-----KKKHVLSGTL---GVPEHTY--
CD2 d1      ADCRDSGTVWGALGHGINLNIPNFQMTD--DIDEVRWERGS-----TLVAEFKRKMKPFLKSG----
                                 *

                 D            E               F                  G
Ig lambda   PARFSGSLI----GNKAALTITGAQTEDEAIYFCALWYSN-----HWVFGGGTKLTVL
Ig kappa    PSRFSGTGS----GTDFTFTISSLQPEDIATYYCQQFDNL-----PLTFGGGTKVDFK
IgG heavy   RSRVTMLVDTS--KNQFSLRLSSVTAADTAVYYCARNLIAG---CIDVWGQGSLVTVS
TcR beta    EDRFSAKMPNA---SFSTLKIQPSEPRDSAVYFCASSFSTCSANYGYTFGSGTRLTVV
TcR alpha   EGRFTIHLNKA--SLHFSLHIRDSQPSDSALYLCAVTLYGG-SGNKLIFGTGTLLSVK
CD8 beta    GERVKKNMTLSFNSTLPFLKIMDVKPEDSGFYFCAMVGSP-----MVVFGTGTKLTVV
CD8 alpha   PNLFSARKE----NNKYILTLSKFSTKNQGYYFCSITSNS-----VMYFSPLVPVFQK
CD4 d1      NDRADSRRSLWD-QGNFPLIIKNLKIEDSDTYICEVE-----------DQKEEVQLLVF
Thy-1       RSRVNLFSDR----FIKVLTLANFTTKDEGDYMCELRVSGQNPTSSNKTINVIRDKLV
CD2 d1      --AFEILA--------NGDLKIKNLTRDDSGTYNVTYSTNG-----TRILDKALDLRIL
                                              *
```

Figure 11. Immunoglobulin C1-set domains

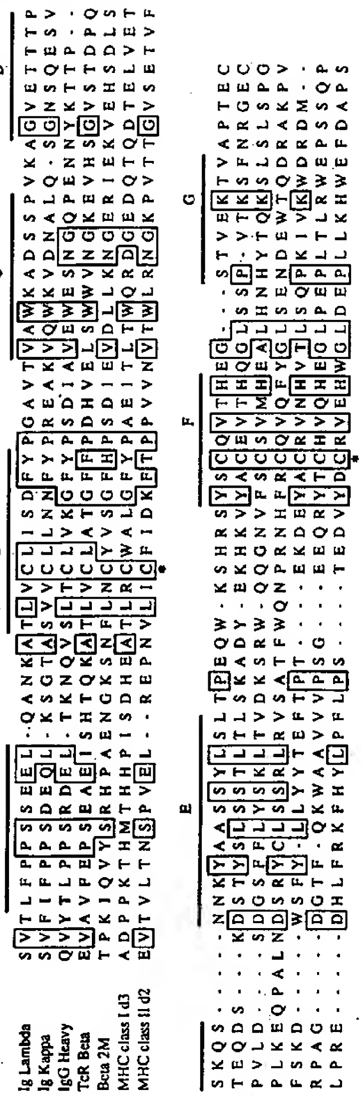


Figure 12. Immunoglobulin C2-set domains

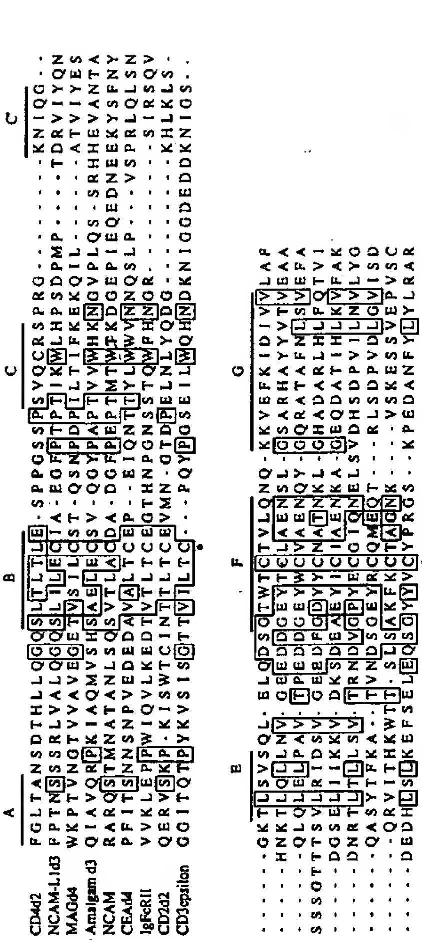


Figure 11. (opposite) *Immunog*
or more sequences are boxed.
the sequences. The asterisks m
marked on the domain organiz
sequences of the following pro
otherwise indicated and the d
given in brackets. Ig Lambda,
Kappa, human Ig κ chain C re
region (P01857, 230–329); TcR
2M, human β 2-microglobulin
MHC Class I HLA α chain pre
II d2, human MHC Class II D
113–209).

Figure 12. (opposite) *Immunog*
or more sequences are boxed.
the sequences. The asterisks r
marked on the domain organi
sequences of the following pr
otherwise indicated and the a
given in brackets. CD4d2, hu
NCAM-L1d3, mouse neural c
(P11627, 243–331); MAGd4, r
4 (P07722, 327–412); Amalga
3 (P15364, 231–327); NCAM,
(P13590, 203–295); CEA d4, h
(P06731, 321–410); IgFcRII, n
37–116); CD2d2, human CD2
human CD3ε precursor (P077

domains, β2-microglobulin [12]
and 2 [8,9] and CD8α [35]. Th
determined by NMR [10]. Thes
by sequence similarities over
with distinct folding patterns
fold consists of a sandwich o
of 5–10 amino acids with a
but not all domains. The seq
in-pointing residues in the β
connect the strands and the
core of the fold is made u
positioning of these is show
vary considerably in length
being the archetype for the I
domains forms an addition
connection between these f
in antibody and TcR V-doma

Figure 11. (opposite) *Immunoglobulin C1-set domains. Residues identical in four or more sequences are boxed. The positions of the β strands are indicated above the sequences. The asterisks mark the positions of the conserved residues that are marked on the domain organization figures in the entries in Section II. The sequences of the following proteins are from the Swissprot database unless otherwise indicated and the database accession number and residue numbers are given in brackets. Ig Lambda, human Ig λ chain C region (P01842, 7–104); Ig Kappa, human Ig κ chain C region (P01834, 6–106); IgG Heavy, human Ig γ-1 C region (P01857, 230–329); TcR Beta, human TcR β chain (P01850, 10–117); Beta 2M, human β 2-microglobulin precursor (P01884, 24–119); MHC class I d3, human MHC Class I HLA α chain precursor domain 3 (PIR:A02189, 203–301); MHC class II d2, human MHC Class II DR α chain precursor domain 2 (PIR:A02206, 113–209).*

Figure 12. (opposite) *Immunoglobulin C2-set domains. Residues identical in four or more sequences are boxed. The positions of the β strands are indicated above the sequences. The asterisks mark the positions of the conserved residues that are marked on the domain organization figures in the entries in Section II. The sequences of the following proteins are from the Swissprot database unless otherwise indicated and the database accession number and residue numbers are given in brackets. CD4d2, human CD4 precursor domain 2 (P01730, 123–204); NCAM-L1d3, mouse neural cell adhesion molecule L1 precursor domain 3 (P11627, 243–331); MAGd4, rat myelin associated glycoprotein precursor domain 4 (P07722, 327–412); Amalgam d3,* Drosophila *amalgam protein precursor domain 3 (P15364, 231–327); NCAM, chicken neural cell adhesion molecule precursor (P13590, 203–295); CEA d4, human carcinoembryonic antigen precursor domain 4 (P06731, 321–410); IgFcRII, mouse IgG FcRII precursor domain 1 (P08101, 37–116); CD2d2, human CD2 precursor domain 2 (P06729, 127–203); CD3 epsilon human CD3ε precursor (P07766, 29–117).*

---

domains, β2-microglobulin [12], MHC Class I antigen α3 domain [34], CD4 domains and 2 [8,9] and CD8α [35]. The structure of domain 1 of CD2 has recently bee determined by NMR [10]. These structures show that the IgSF domains characterize by sequence similarities over about 100 amino acids correspond to structural uni with distinct folding patterns referred to as the Ig-fold [reviewed in ref. 12]. The I fold consists of a sandwich of two β sheets, each consisting of antiparallel β stran of 5–10 amino acids with a conserved disulphide between the two sheets in mo but not all domains. The sequence similarities are mainly found at the positions in-pointing residues in the β strands with considerable differences in the loops th connect the strands and the out-pointing residues on the faces of the β sheets. T core of the fold is made up of three β strands labelled ABE and GFC and t positioning of these is shown in the various folds illustrated in Fig. 9. The fol vary considerably in length in the middle of the sequence with Ig V-domain fol being the archetype for the longer fold. The extra sequence in comparison with domains forms an additional pair of β strands (C' and C" in Fig. 9) and t connection between these forms the second complementarity determining regi in antibody and TcR V-domains.

Figure 13. *Integrin α chains. Residues identical in 3 out of 4 of the sequences are boxed. The sequences of the following proteins are from the Swissprot database and the database accession number and residue numbers are given in brackets. CD49a, rat integrin α1 precursor (P18614, 25–1172); CD49b, human integrin α2 precursor (P17301, 26–1161); CD51, vitronectin receptor integrin α V precursor (P06756, 27–1023); CD49d, integrin α4 precursor (P13612, 36–1013). The extracellular and transmembrane regions are shown.*

```
CD49A  FCVSFNVDVKNSMSFSGPVEDMFGYTVQQYENEEG--KWVLIGSPLVGQPKA----RTGDVYKCPVGRERAMPCVKLDLPV
CD49B  CCLAYNVGLPEAKIFSGPSSEQFGYAVQQFINPKG--NWLLVGSPWSGFPEN----RMGDVYKCPVDLSTA-TCEKLNLQT
CD51   LCRAFNLDVDSPAEYSGPEGSYFGFAVDFFVPSASSRMFLLVGAPKAN-TTQPGIVEGGQVLKC--DWSSTRRCQPIEFDA
CD49D  TGRPYNVDTESALLYQGPHNTLFGYSV--VLHSHGANRWLLVGAPTANWLANASVINPGAIYRCRIGKNPGQTCEQLQLGS

NTSIP----NVTEIKENMTFGSTLVTNP--NGGFLACGPLYAYRCGHLHY------TTGICSDVSPTFQVVNSFAP-VQECSTQ-
STSIP----NVTEMKTNMSLGLILTRNMG-TGGFLTCGPLWAQQCGNQYY------TTGVCSDISPDFQLSASFSPATQPCPSL-
TGNRDYAKDDPLEFKSHQWFGASV---RSKQDKILACAPLYHWRTEM----KQEREPVGTCFLQDGTKTVEY------APCRSQD
PNGEPCGK-TCLEERDNQWLGVTLSRQPGENGSIVTCG---HRWKNIFYIKNENKLPTGGCYGVPPDLRTELSKRI--APCYQDY

LDIVIVLDGSN - 196 Amino acids --SQTGFSAHYS--QDWVMLGAVGAYDWNGTVVMQKANQMV---------IPHNTTFQTEPAKMNE
IDVVVVCDESN - 195 Amino acids --SQVGFSADYSSQNDILMLGAVGAFGWSGTIVQKTSHGHLI--------FPKQAFDQILQDRNH-
IDADGQGF----- (I Domain) ---CQGGFSIDFT-KADRVLLGGPGSFYWQGQLISDQVAEIVSKYDPNVYSIKYN--NQLATRTAQA
VKKFGENFAS-----------CQAGISSFYT--KDLIVMGAPGSSYWTGSLF-------------VYNITTNKYKAFLDKQNQV

PL-ASYLGYTVNSATIPGD--VLYIAGQPRYNHT-GQVVIYKMEDGNINILQTLGGEQIGSYFGSVLTTIDIDKDSYTDLLLVGA
---SSYLGYSVAAISTGES--THFVAGAPRANYTGQIVLYSVNENGNITVIQAHRGDQIGSYFGSVLCSVDVDKDTITDVLLVGA
IFDDSYLGYSVAVGDFNGDGIDDFVSGVPRAARTLGMVYIY--DGKNMSSLYNFTGEQMAAYFGFSVAATDINGDDYAD-VFIGA
KF-GSYLGYSVGAGHFRSQHTTEVVGGAPQHEQ-IGKAYIFSIDEKELNILHEMKGKKLGSYFGASVCAVDLNADGFSD-LLVGA

PMYM----GTEKEEQGKVYVYAVNQTRFEYQMSLEPIRQTCCSSLKDNSCTKENKNEPCGARFGTAIAAVKDLNVDGFNDVVIGA
PMYM----SDLKKEEGRVYLFTIKKGILGQHQFL-------------------EGPEGIENTRFGSAIAALSDINMDGFNDVIVGS
PLFMDRGSDGKLQEVGQVSVSLQRASGDF------------------QTTKLNGFEVF-ARFGSAIAPLGDLDQDGFNDIAIAA
PM------QSTIREEGRVFVYINSGSGAV-----------------MNAMETNLVGSDKYAARFGESIVNLGDIDNDGFEDVAIGA

PL-EDDHAGAVYIYHGSGKTIREAYAQRIPSGGDG--KTLKFFGQSIHGEMDLNGDGLTDVTIGGLG--GAALFWARDVAVVKVT
PL-ENQNSGAVYIYNGHQGTIRTKYSQKILGSDGAFRSHLQYFGRSLDGYGDLNGDSITDVSIGAFG--QVVQLWSQSIADVAIE
PYGGEDKKGIVYIFNGRSTGLNAVPSQILEGQWAA-RSMPPSFGYSMKGATDIDKNGYPDLIVGAFGVDRAILYRARPVITVNAG
P-QEDDLQGAIYIYNGRADGISSTFSQRIEGLQIS-KSLSM-FGQSISGQIDADNNGYVDVAVGAFRSDSAVLLRTRPVVIVDAS

MNFEPNKVNIQKKNCRVEGK--ETVCINATMCFHVKLKSKEDSIYEADLQYRVTLD----SLRQISRSFFSGTQERKIQRNITVR
ASFTPEKITLVNKNAQIILK---------LCFSAKFRPTKQN-NQVAIVYNITLDADGFSSRVTSRGLFKENNERCLQKNMVVN
LEVYPSILNQDNKTCSLPGTALKVSCFNVRFCLKADGKGV--LPRKLNFQVELLLDKLKQKGAIRRALFLYSRSPSHSKNMTI--
LS-HPESVNRTKFDC--VENGWPSVCIDLTLCFSYKGKEV---PGYIVLFYNMSLD-VNRKAESPPRFYFSSNGTSDVITGSIQV
```



```
ESE-----CIRHSFYMLDKHDFQD---SVRVTLDFNLTD-----------PENGPVLDDALPNSVHEHIPFAKDCGNKERCISDL
QAQS----CPEHIIYIQEPSDVVN---SLDLRVDISLEN----------PGTSPALEAYSETAKVFSIPFHKDCGEDGLCISDL
-SRGGLMQCEELIAYLRDESEFRDKLTPITIFMEYRLD----YRTAADTTGLQPILNQFTPANISRQAHILLDCGEDNVCKPKL
SSREA--NCRTHQAFMRK--DVRDILTPIQIEAAYHLGPHVISKRSTEEFPPLQPILQQKKEKDIMKKTINFARFCAHENCSADL

TLNV-----STTEKSLLIVKSQHDKFNVSLTVKNKGDSAYNTRTVVQHSPNLIFSGI----EEIQKDSCESNQ-  -NITCRVG
VLDV--RQIPAAQEQPFIVSNQNKRLTFSVTLKNKRESAYNTGIVVDFSENLFFASF-SLPVDGTEVTCQVAASQK--SVACDVG
EVSV-----DSDQKKIYIGDDNPLTLIVKAQ--NQGEGAYEAELIVSIPLQADFIGVVRNNEALARLSCAFKTENQTRQVVCDLG
QVSAKIGFLKPHENKTYLAVGSMKTLMLNVSLFNAGDDAYETTLHVKLPVGLYFIKILELEE--KQINCEVTDNSGVVQLDCSIG
```

```
ESE-----CIRHSFYMLDKHDFQD---SVRVTLDFNLTD-----------PENGPVLDDALPNSVHEHIPFAKDCGNKERCISDL
QAQS----CPEHIIYIQEPSDVVN---SLDLRVDISLEN-----------PGTSPALEAYSETAKVFSIPFHKDCGEDGLCISDL
-SRGGLMQCEELIAYLRDESEFRDKLTPITIFMEYRLLD-----YRTAADTTGLQPILNQFTPANISRQAHILLDCGEDNVCKPKL
SSREA--NCRTHQAFMRK--DVRDILTPIQIEAAYHLGPHVISKRSTEEFPPLQPILQQKKEKDIMKKTINFARFCAHENCSADL

TLNV-----STTEKSLLIVKSQHDKFNVSLTVKNKGDSAYNTRTVVQHSPNLIFSGI----EEIQKDSCESNQ-   -NITCRVG
VLDV--RQIPAAQEQPFIVSNQNKRLTFSVTLKNKRESAYNTGIVVDFSENLFFASF-SLPVDGTEVTCQVAASQK--SVACDVG
EVSV-----DSDQKKIYIGDDNPLTLIVKAQ--NQGEGAYEAELIVSIPLQADFIGVVRNNEALARLSCAFKTENQTRQVVCDLG
QVSAKIGFLKPHENKTYLAVGSMKTLMLNVSLFNAGDDAYETTLHVKLPVGLYFIKILELEE--KQINCEVTDNSGVVQLDCSIG

YPFLRAGETVTFKIIFQFN-TSHLSENAIIHLSATSDS-EEPLESLNDNEVNISIPVKYEVGLQFYSSASEHHISVAANETIPEF
YPALKREQQVTFTINFDFN-LQNLQNQASLSFQALSESQEE---NKADNLVNLKIPLLYDAEIHLTRSTNINFYEISSDGNVPSI
NP-MKAGTQLLAGLRFSVHQQSEMDTSVKFDLQIQSSNLFDKVSPVVSHKVDLAVLAAVEIR---GVSSPDHIFLPIPNWEHKEN
YIYVDHLSRIDISFLLDVSSLSRAEEDLSITVHATCENEEE-MDNLKHSRVTVAIPLKYEVK------LTVHGFVNPTSFVYGSN

INSTEDIGN--EINVFYTIRKRGHFPMPELQLSISFPN-LTADGYPVLYPIGWS-SSDNVNCRPRSLEDPFGINSGKKMTISKSE
VHSFEDVGP---KFIFSLKVTTGSVPVSMATVIIHIPQ-YTKEKNPLMYLTGVQ-TDKAGDISCNADINPLKIGQTSSSVSFKSE
PETEEDVGPV--VQHIYELRNNGPSSFSKAMLHLQWPYKYNNNT--LLYILHYD-IDGPMNCTSDMEINPLRIKISSLQTTEKND
DENEPETCMVEKMNLTFHVINTGNSMAPNVSVEIMVPNSFSPQTDKLFNILDVQTTTGECHFENYQRVCALEQQKSAMQTLKGIV

VLKRGTIQD------------------CSS-TCGVATITCSLLPSDLSQ---VNVSLLLWKPTFIRAHFSSLNLTLRG------
NFRHTKELN------------------CRTASCS--NVTCWLKDVHMKGEYFVNVTTRIWNGTFASSTFQTVQLTAAA------
TVAGQGERDHLITKRDLALSEGDIHTLGCGVAQCL--KIVCQVGRLDRGKSAILYVKSLLWTETFMNKENQNHSYSLKSSASFNV
RFLSKTDKRLLYCIK-------------ADPHCL--NFLCNFGKMESGKEASVHIQLEGRPSILEMDETSALKFEIRATG----

-ELKSENSSLTLSSSNRKRELAIQISKDGLPGRVPLWVIIL---SAFAGLLLLMLLILALWKIGFFKR
-EINTYNPEIYVIEDN-TVTIPLMIMKPDEKAEVPTGVIIG---SIIAGILLLALVAILWKLGFFKR    cytoplasmic domain
IEFPYKNLPIEDITNSTLVTTNVTWGIQPAPMPVPVWVIIL---AVLAGLLLLAVLVFVMYRMGFFKR
--FPEPNPRVIELNKDENV-AHVLLEGLHHQRPKRYFTIVIISSSLLLGLIVLLLISYVMWKAGFFKR
```

In the IgSF there are limited sequence patterns in β strands B, C, E, and F that are common across the superfamily (Figs 10-12) and other limited patterns that allow a subdivision of the domains. Ig and TcR V-domains have a characteristic pattern in the region leading into β strand F of Asp-X-Gly/Ala-X-Tyr-X-Cys. The receptor C-domains have a characteristic pattern between β strands B and C of GlyPheTyrPro and another on the COOH-terminal side of β strand F of Cys-X-Val-X-His. The Ig, TcR and MHC antigen C-type domains all share the same types of sequence patterns and are referred to as the C1 set within the IgSF. With the sequencing of various cell surface molecules a third category of domains became evident, namely domains of length similar to C-domains but with some of the sequence patterns of V-domains. These domains are referred to as the C2 set [2]. They have V-type patterns in the β strand E to F region, a pattern of Pro-X-Pro is relatively common between β strands B and C and the pattern Cys-X-Ala-X-Asn is common after β strand F. CD4 domain 2 is a C2-set sequence and its structure is classified in terms of sheet assignments labelled as ABE/GFCC'. This is in comparison to ABED/GFC for C1-set sequences and ABED/GFCC'C" for V-set sequences. That is, for C2-set sequences the middle β strand may be generally in line with the GFC β sheet rather than the ABE sheet as is the case with antibody C-domains. The points about conserved patterns and the positioning in β sheets are made evident by comparing the sequence alignments in Figs 10–12 with the folding patterns for the domains in Fig. 9. The sequence alignments are discussed in more detail in ref. 2.

Integrin superfamily (Figs 13 and 14)

The integrins are a large family of related proteins that all share a heterodimeric structure with α and β chains that both traverse the lipid bilayer. There are at least 20 α and eight β chains which can be found in various but not all possible combinations. Sequence similarities are seen within the α and β chain across all the integrin types (Figs 13 and 14). The integrins are known to be involved in cell interactions and include receptors for the extracellular matrix proteins fibronectin and vitronectin and for cell surface molecules ICAM-1 and ICAM-2. The integrins have been extensively reviewed elsewhere including a companion volume in this FactsBook series [36]. They are expressed on many different cell types; the CD11/CD18 family and the CD49 very late activation antigen family (VLA) are expressed mainly on leucocytes. The sequence similarities in this family are described in more detail in refs 37–39.

This family of related proteins does not contain other domain types apart from the β4 integrin that contains two FN type III SF domains in the cytoplasmic region (see ref. 33 and section on fibronectin type III SF domains).

Figure 14. (opposite) *Integrin β chains. Residues identical in 3 out of 3 of the sequences are boxed. The sequences of the following proteins are from the Swissprot database and the database accession number and residue numbers are given in brackets. Beta 1, human fibronectin receptor, integrin β_1 precursor (P05556, 26–752); Beta 2, human integrin β_2 (CD18) precursor (P05107, 024–724); Beta 3, human integrin β_3 (CD61) precursor (P05106, 30–742). The extracellular and transmembrane regions are shown.*

that are
t allow a
attern in
eptor C-
ieTyrPro
. The Ig,
sequence
ncing of
, namely
tterns of
e V-type
common
a after β
in terms
ED/GFC
or C2-set
et rather
ts about
mparing
mains in

odimeric
e at least
possible
cross all
d in cell
ronectin
integrins
e in this
pes; the
VLA) are
mily are

art from
ic region

3 of the
rom the
ibers are
precursor
24–724);
acellular

Figure 14. Integrin β chains

```
Beta 1  RCLKANAKSCGECIQAGPNCGWCTNSTF-LQEGMPTSARCDDLEALKKKGCPPDDIENPRGSKDIKKNKNVT
Beta 2  ECTKFKVSSCRECIESGPGCTWCQKLNF-TGPGDPDSIRCDTRPQLLMRGCAADDIMDPTSLAETQEDHNGG
Beta 3  ICTTRGVSSCQQCLAVSPMCAWCSDEALPLGSP-----RCDLKENLLKDNCAPESIEFPVSEARVLEDRPLS

NRSKGTAEKLKPEDIHQIQPQQLVLRLRSGEPQTFTLKFKRAEDYPIDLYYLMDLSYSMKDDLENVKSLGTDLMNE
QK-------------QLSPQKVTLYLRPGQAAAFNVTFRRAKGYPIDLYYLMDLSYSMLDDLRNVKKLGGDLLRA
DKGSGDSS-----QVTQVSPQRIALRLRPDDSKNFSIQVRQVEDYPVDIYYLMDLSYSMKDDLWSIQNLGTKLATQ

MRRITSDFRIGFGSFVEKTVMPYISTTP-AKLRNPCTS-EQNCTTPFSYKNVLSLTNKGEVFNELVGKQRISGNLD
LNEITESGRIGFGSFVDKTVLPFVNTHP-DKLRNPCPNKEKECQPPFAFRHVLKLTNNSNQFQTEVGKQLISGNLD
MRKLTSNLRIGFGAFVDKPVSPYMYISPPEALENPCYDMKTTCLPMFGYKHVLTLTDQVTRFNEEVKKQSVSRNRD

SPEGGFDAIMQVAVCGSLIGWRNVT-RLLVFSTDAGFHFAGDGKLGGIVLPNDGQCHLE-NNMYTMSHYYDYPSIA
APEGGLDAMMQVAACPEEIGWRNVT-RLLVFATDDGFHFAGDGKLGAILTPNDGRCHLE-DNLYKRSNEFDYPSVG
APEGGFDAIMQATVCDEKIGWRNDASHLLVFTTDAKTHIALDGRLAGIVQPNDGQCHVGSDNHYSASTTMDYPSLG

HLVQKLSENNIQTIFAVTEEFQPVYKELKNLIPKSAVGTLSANSSNVIQLIIDAYNSLSSEVILENGKLSEGVTIS
QLAHKLAENNIQPIFAVTSRMVKTYEKLTEIIPKSAVGELSEDSSNVVHLIKNAYNKLSSRVFLDHNALPDTLKVT
LMTEKLSQKNINLIFAVTENVVNLYQNYSELIPGTTVGVLSMDSSNVLQLIVDAYGKIRSKVELEVRDLPEELSLS

YKSYCKNGVNGTGENGRKCSNISIGDEVQFEISITSNKCPKKDSDSFKIRPLGFTEEVEVILQYICECECQSEGIP
YDSFCSNGVTHRNQPRGDCDGVQINVPITFQVKVTATECIQEQ--SFVIRALGFTDIVTVQV--LPQCECRCRDQS
FNATCLNNEVIPGL--KSCMGLKIGDTVSFSIEAKVRGCPQEKEKSFTIKPVGFKDSLIVQVTFDCDCACQAQAEP

ES-PKCHEGNGTFECGACRCNEGRVGRHCECSTDEVNSEDMDAYCRKENSSEICSNNGECVCGQCVCRKRDNTNEI
RDRSLCH-GKGFLECGICRCDTGYIGKNCECQTQGRSSQELEGSCRKDNNSIICSGLGDCVCGQCLCHTSDVPGKL
NS-HRCNNGNGTFECGVCRCGPGWLGSQCECSEEDYRPSQQDE-CSPREGQPVCSQRGECLCGQCVCHSSDFGK--

YSGKFCECDNFNCDRSNGLICG--GNGVCKCRVCECNPNYTGSACDCSLDTSTCEASNGQICNGRGICECGVCKCT
IYGQYCECDTINCERYNGQVCGGPGRGLCFCGKCRCHPGFEGSACQCERTTEGCLNPRRVECSGRGRCRCNVCEC-
ITGKYCECDDFSCVRYKGEMCS--GHGQCSCGDCLCDSDWTGYYCNCTTRTDTCMSSNGLLCSGRGKCECGSCVCI

DPKFQGQTCEMCQ[illegible]---------NPVKDR[illegible]
HSGYQLPLCQECPGCPSPCGKYISCAECLKFEKGPF--GKNCSAACPGLQLSN-----------TGKDAVNCTYKN
QPGSYGDTCEKCPTCPDACTFKKECVECKKFDREPYMTENTCNRYCRDE-IESVKELKD-----

VDDCWFYFTYS-VNGNNEVMVHVVENPECPTGPDIIPIVAGVVAGIVLIGLALLLIWK
SEGCWVAYTLEQQDGMDRYLIYVDESRECVAGPNIAAIVGGTVAGIVLIGILLLVIWK
EDDCVVRFQYYEDSS-GKSILYVVEEPECPKGPDILVVLLSVMGAILLIGLAALLIWK
```

cytoplasmic domain

Figure 15. *Lectin C-type superfamily domains. Residues identical in 4 out of 7 of the sequences are boxed. The positions of the β strands (β), α helixes (α) and loops (L) determined for the structure of the rat mannose binding protein [49] are shown above the sequences. The asterisks mark the positions of the conserved residues that are marked on the domain organization figures in the entries in Section II. The sequences of the following proteins are from the Swissprot database and the database accession number and residue numbers are given in brackets. ManBP, rat mannose binding protein A precursor (P19999, 117–238); ManR, human mannose receptor precursor (P2897, 362–490); KUCR, Kupffer cell carbohydrate binding receptor (P10716; 412–540); LECH, rat hepatic lectin-1 or asialoglyoprotein receptor-1 (P02706; 152–279); CD23, low affinity IgE receptor (P06734; 163–286); PGCA, cartilage specific proteoglycan core (P07897; 1914–2038); E-Sel, E-selectin or ELAM-1 precursor (P16581; 22–142); CD62, CD62 or granule membrane protein 140 or P-selectin precursor (P16109; 42–162); L-Sel, human leucocyte adhesion molecule or L-selectin precursor (P14151; 39–159).*

```
                                 β1            α1                       α2               β2
ManBP  AFSMQKKSGKKFFVTNHERM-PFSKVKALCSELRGTVAIPRNAEENKAIQEVAK----TSAFLGIT
ManR   CPSQWWPYAGHCYKIHRDEKKIQRDALTTCRKEGGDLTSIHTIEELDFIISQLGYEPNDELWIGLN
KUCR   IMQDWKYFNGKFYYFSRDKK-SWHEAENFCVSQGAHLASVTSQEEQAFLVQITN---AVDHWIGLT
LECH   CPINWVEYEGSCYWFSSSVK-PWTEADKYCQLENAHLVVVTSWEEQRFVQQHMG---PLNTWIGLT
CD23   CPEKWINFQRKCYYFGKGTK-QWVHARYACDDMEGQLVSIHSPEEQDFLTKHA---SHTGSWIGLR
PGCA   CEEGWTKFQGHCYRHFPDRE-TWVDAERRCREQQSHLSSIVTPEEQEFVNKNA----QDYQWIGLN
E-Sel             WSYNTSTEAM-TYDEASAYCQQRYTHLVAIQNKEEIEYLNSILSY-SPSYYWIGIR
CD62              WTYHYSTKAY-SWNISRKYCQNRYTDLVAIQNKNEIDYLNKVLPY-YSSYYWIGIR
L-Sel             WTYHYSEKPM-NWQRARRFCRDNYTDLVAIQNKAEIEYLEKTLPF-SRSYYWIGIR
                                     *

 L1      L2            L3          L4         β3              β4          β5
DEVTEGQFMYVTGGRLT----YSNWKKDEPNDHGS-----GEDCVTIVDN-----GLWNDISCQA-SHTAVCEFPA
DIKIQMYFEWSDGTPVTFTKWLRGEPSHENNRQ--------EDCVVMKG----KDGYWADRGCEW-PLGYICKMKS
DQGTEGNWRWVDGTPFDYVQSRRFWRKGQPDNW-RHGNGEREDCVHL-------QRMWNDMACGT-AYNWVCKSTD
DQ--NGPWKWVDGTDY--ETGFKNWRPGQPDDWYGHGLGGGEDCAHF-----TTDGHWNDDVCRR-PYRWVCETEL
NLDLKGEFIWVDGSHV----DYSNWAPGEPTSR-----SQGEDCVMM-----RGSGRWNDAFCDRKLGAWVCDRLA
DRTIEGDFRWSDGHSL----QFEKWRPNQPDNFF----ATGEDCVVM---IWHERGEWNDVPCNY-QLPFTCKKGT
K--VNNVWVWV-GTQKPLTEEAKNWAPGEPNNR-----QKDEDCVEIYIKREKDVGMWNDERCSK-KKLALCYTAA
K--NNKTWTWV-GTKKALTNEAENWADNEPNNK-----RNNEDCVEIYIKSPSAPGKWNDEHCLK-KKHALCYTAS
K--IGGIWTWV-GTNKSLTEEAENWGDGEPNNK-----KNKEDCVEIYIKRNKDAGKWNDDACHK-LKAALCYTAS
                                                                   *
```

Lectin C-type superfamily (Figs 15 and 16)
This family of lectin domains are termed C-type because some members have been shown to bind carbohydrate in a Ca^{2+} dependent reaction [40]. This domain has been found in a number of lectins such as the Kupffer cell fucose/galactose receptor, hepatocyte galactose receptor, mannose binding protein from plasma, and galactose binding proteins in two invertebrate species, the flesh fly and sea urchin [41–43]. Lectin C-typeSF domains are found in a number of proteins not originally known to bind carbohydrate, such as the proteoglycan core protein [44], an endothelial leucocyte adhesion molecule (E-selection), and leucocyte cell surface antigens such as L-selectin and the low affinity Fc receptor for IgE (CD23). In some cases carbohydrate binding for the lectin C-typeSF domain has been established, e.g. cartilage proteoglycan core protein [45].

Two groups of lectin C-type domains can be distinguished. The L-selectin has the lectin C-typeSF domain plus about 10 residues of the signal sequence contained completely within one exon with phase 1 intron boundaries [46]. In cases other than the selectins, the lectin domain is usually found spread over three exons which also include the COOH-terminus of the protein and the 3' untranslated sequence [47]. In the genetic region encoding the NH_2-terminal side of the exon there is a phase 1 intron boundary and thus if this exon was inserted into an intron of another gene it might function to produce a new protein with the lectin domain at the COOH-terminus. If the insertion occurred after a hydrophobic region in a Type 1 membrane protein the lectin exon would be cytoplasmic, and thus presumably of little functional relevance. Conversely if it were inserted in a Type II protein it would form a new extracellular COOH-terminal region with carbohydrate binding properties. The recent cDNA sequence of the macrophage mannose receptor [48] is the first example of a protein containing multiple lectin repeats, with eight in all. There is no information as yet on its genomic structure.

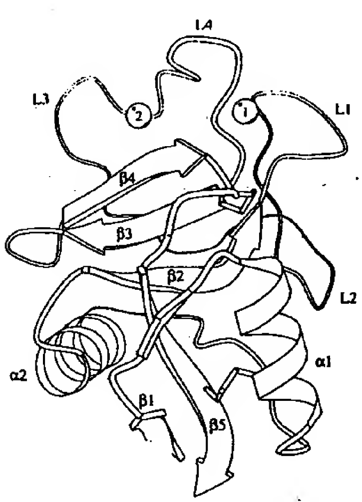


Figure 16. *The folding pattern of a lectin C-typeSF domain. Ribbon diagram of the lectin C-typeSF domain from the rat mannose binding protein [49]. The β strands are shown as broad arrows pointing from the amino to carboxy direction, α helices as coiled ribbons and the connecting loops as thinner lines. The labelling of the β strands (β1–5), α helices (α1–2) and loops (L1–4) corresponds to that in the sequence alignments in Fig. 15. The numbers 1 and 2 refer to the position of the two holmium ions that are known to stabilize this region that contains a high proportion of nonregular secondary structure [49].*

Figure 17. *Lectin S-type superfamily domains. Residues identical in 3 out of 5 of the sequences are boxed. The complete sequences are given for galactose binding lectins from four species together with the region showing sequence similarity in the Mac-2 antigens from human and rat. The sequences of the following proteins are from the Swissprot database and the database accession number and residue numbers are given in brackets. Mac2-Human, human Mac-2 antigen precursor (P17931, 111–248); Mac2-Rat, rat Mac-2 antigen precursor (P08699, 123–260); G-lectin-Rat, rat β galactoside binding lectin (P11762); G-lectin-Chick, chicken β galactoside binding lectin (P07583); G-lectin-Human, human β galactoside binding lectin (P09382); G-lectin-eel, electric eel β galactoside binding lectin (P08520).*

```
Mac2-Human      AGPLIVPYNLPL-PGGVVPRMLITILGTVKPNANRIALDFQR
Mac2-Rat        TGPLTVPYDMPL-PGGVMPRMLITIIGTVKPNANSITLNFKK
G-lectin-Rat    AC-GLVASNLNLKPGEC------LKVRGELAPDAKSFVLNLGK
G-lectin-Chick  SCQGPVCTNLGLKPGQR------LTVKGIIAPNAKSFVMNLGK
G-lectin-Human  AC-GLVASNLNLKPGEC------LRVRGEVAPDAKSFVLNLGK
G-lectin-Eel    SMNGVVDERMSFKAGQN------LTVKGVPSIDSTNFAINVGN

       *                                             *
G-NDVAFHFNPRFN-ENNRRVIVCNTKLDNNWGREERQSVFPFESGKPFK
G-NDIAFHFNPRFN-ENNRRVIVCNTKQDNNWGREERQSAFPFESGKPFK
DSNNLCLHFNPRFNAHGDANTIVCNSKDDGTWGTEQRETAFPFQPGSITE
DSTHLGLHFNPRFDAHGDVNLIVCNSKKMEEWGTEQRETVFPFQKGAPIE
DSNNLCLHFNPRFNAHGDANTIVCNSKDGGAWGTEQREAVFPFQPGSVAE
SAEDLALHINPRFDAHGDQQAVVVNSFQGGNWGTEQREGGFPFKQGEDFK

IQVLVEPDHFKVAVNDAHLLQYNHRVKKLNEISKLGISGDIDLTSASYT
IQVLVEADHFKVAVNDVHLLQYNHRMKNLREISQLGIIGDITLTSASHA
VCITFDQADLTIKLPDGHEFKFPNRL-NMEAINYMAADGDFKIKCVAFE
ITFSINPSDLTVHLP-GHQFSFPNRL-GLSVFDYFDTHGDFTLRSVSWE
VCITFDQANLTVKLPDGYEFKFPNRL-NLEAINYMAADGDFKIKCVAFD
IQITFNSEEFRIILPDGSEIHFPNN---------RYMHFEGEARIYSIEIK
```

As well as the differences in intron/exon organization there ar sequence patterns that differ between these two groups of lectin C-typeSF don ins, as is evident in the sequence alignments shown in Fig. 15. There is a charac ristic Trp residue at the NH_2-terminus of the selectins E-selectin, CD62 (P-sele n) and L-selectin, whilst in the other group there is a longer patch of sequence the NH_2-terminus that shows a conserved sequence pattern of Cys-X-X-X-X-Trp.

The structure of a lectin C-typeSF domain from a rat mannose binc g protein has recently been determined by X-ray crystallography [49] and the fold shown in Fig. 16. The structure is unusual in that it contains two regions, or of which contains non-regular secondary structure stabilized by two holmium ns in the crystal structure. The other contains both β sheet and α helix; this is unu al as most domains for the other cell surface molecules consist solely of β structu .

Lectin S-type superfamily (Fig. 17)
Galactoside binding proteins have been sequenced from several specie nd shown to contain a sequence pattern different from that of the lectin C-typeSF. hese have been termed S-type because the first examples contained free acc sible thiol groups [40]. These are found both intracellularly and extracellularly and a egion with strong sequence similarity is found in the Mac-2 leucocyte antig (Fig. 17). However, analysis of protein produced by recombinant DNA techniqu shows no requirement for a reducing environment for lectin activity and no acc sible thiol groups [50].

```
         *                                         *
CD42a    L L L A N N S L Q S V P P G A F D H L P Q L Q
CD42b    L V L T G N N L T A L P P G L L D A L P A L R
A2g-1    L D L S G N R L R K L P P G L L A N F T L L R
A2g-2    L D L G E N Q L E T L P P D L L R G P L Q L E
A2g-3    L H L E G N K L Q V L G K D L L L P Q P D L R
PG-1     L D L Q N N K I T E I K D G D F K N L K N L H
PG-2     L I L V N N K I S K V S P G A F T P L V K L E
PG-3     L Y L S K N Q L K E L P E K M P K T L Q E L R
Chao-1   L I L P Q N D L V E I P S K S L R H L Q K L R
Chao-2   L D L G Y N H I T H I Q H D S F R G L E D S L
CYCL-1   L E L Q R N F I R K V P N S I M K - L S N L T
CYCL-2   L N L Q C N E L E S L P A G F V E - L K N L Q
```

Figure 18. *Leucine-rich glycoprotein repeats. Residues identical in six r more sequences are boxed. The asterisks mark the positions of the conserve residues that are marked on the domain organization figures in the entries in ction II. The sequences of the following proteins are from the Swissprot datab e unless otherwise indicated and the database accession number and residue mbers are given in brackets. CD42a, human platelet glycoprotein IX precursor (4770, 56–79); CD42b, human platelet glycoprotein 1B β chain precursor (P1 24, 60–83); A2g-1, A2g-2 and A2g-3, human leucine-rich α2 glycoprotein (LRG) (2750, 134–157, 158–181 and 182–205 respectively); PG-1, PG-2 and PG-3, h nan bone proteoglycan II precursor (P07585, 86–109,110–133 and 134–157 resp tively); Chao-1 and Chao-2,* Drosophila *chaoptin precursor (P12024, 132–155 nd 156–179 respectively); CYCL-1 and CYCL-2, yeast adenylate cyclase (P08678, 58–890 and 891–913 respectively).*

Leucine-rich repeats (LRR) or leucine-rich glycoprotein (LRG) repeats (Fig. 18)
The leucine-rich repeat (LRR) is characterized by a pattern of conserved residues including about 5 or 6 leucines and some other residues in a tightly defined repeat of 24 residues (Fig. 18). It is found both intracellularly and extracellularly in a variety of species including *Drosophila* and yeast and has also been found in the platelet glycoproteins CD42a and CD42b. It often occurs in an array of tandem repeats. For instance there are nine repeats in the leucine-rich glycoprotein where the repeat was first noted, 26 repeats in the yeast adenyl cyclase, 10 in the proteoglycan protein [51] and three in the trkB protein [52]. In some cases some sequence similarity is observed beyond the alignments shown and an alternative alignment may start from the conserved Pro position which is in the centre of the alignment shown. The α chain of the CD42b contains seven LRRs and these, together with all the remaining coding sequence, are encoded by a single exon [53]. Thus, this repeat is not generally coded by single exons in this case. The LRRs have been found in diverse proteins and they have been implicated in the specificity of hormone binding to gonadotropin receptors [54] and in the interaction between yeast adenylate cyclase and RAS proteins [55].

The leucine and other residues in LRRs form an amphipathic sequence which could be involved in protein–protein or protein–lipid interactions. One of the repeats of the *Drosophila* chaoptin protein has been synthesized. This peptide is soluble in aqueous solution but will bind to phospholipid vesicles where it forms predominantly a β structure. It has been suggested that protein segments containing tandem repeats may also form amphipathic β sheets [56].

Link superfamily (Fig. 19)
Two link superfamily domains were originally noted in the link protein that binds hyaluronic acid [57]. This protein also has one IgSF domain. Subsequently, a further four linkSF domains were observed in the proteoglycan core protein that has a chondroitin sulphate binding site. This protein also contains one IgSF domain, a CCPSF domain and a lectin C-typeSF domain [44]. There is also a single linkSF domain in the CD44 antigen which is known to bind to hyaluronate [58, 59].

Low density lipoprotein receptor (LDLR) superfamily (Fig. 20)
The LDL receptor contains seven domains of about 40 amino acids with six conserved cysteine residues that have been called LDLRSF domains [60]. The LDLR also contains three EGFSF domains. LDLRSF domains have also been found in other proteins, notably some complement components such as C6, C9 and factor I. In the LDLR, four of the LDLRSF domains are each encoded by one exon whilst the other three are encoded by a single exon [61]. In the LDLR mutational analysis has indicated that the LDLRSF domains are important in the binding of some lipoproteins but otherwise the function of this domain type is not known [62]. The structure of the LDLSF domain has not been determined to date.

Ly-6 superfamily (Fig. 21)
The Ly-6 antigens are a group of leucocyte antigens first identified in the mouse that consist of 70–80 amino acids containing 10 Cys residues [63, 64]. Southern blot analysis indicates that many Ly-6-related genes are present in the mouse and of these, 10 distinct genes have been identified [65]. The Ly-6 antigens are expressed in non-lymphoid tissues, for example, kidney, as well as on leucocytes. Homologues

Figure 19. *Link superfamily domains. Residues identical in four or more sequences are boxed. The asterisks mark the positions of the conserved residues that are marked on the domain organization figures in the entries in Section II. The sequences of the following proteins are from the Swissprot database and the database accession number and residue numbers are given in brackets. CD44, human CD44 antigen precursor (P16070, 32–123); CORE1, CORE2, CORE3 and CORE4, rat cartilage-specific proteoglycan core protein precursor (P07897, 152–251, 253–353, 486–585 and 587–687 respectively); LINK1 and LINK2, rat proteoglycan link protein (P03994, 143–242 and 244–339 respectively).*

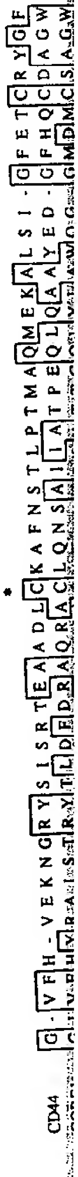


Figure 19. *Link superfamily domains. Residues identical in four or more sequences are boxed. The asterisks mark the positions of the conserved residues that are marked on the domain organization figures in the entries in Section II. The sequences of the following proteins are from the Swissprot database and the database accession number and residue numbers are given in brackets. CD44, human CD44 antigen precursor (P16070, 32–123); CORE1, CORE2, CORE3 and CORE4, rat cartilage-specific proteoglycan core protein precursor (P07897, 152–251, 253–353, 486–585 and 587–687 respectively); LINK1 and LINK2, rat proteoglycan link protein (P03994, 143–242 and 244–339 respectively).*

```
                               *
CD44    G-VFH-VEKNGRYSISRTEAADLCKAFNSTLPTMAQMEKALSI-GFETCRYGF
CORE1   GIVFHYRAISTRYTLDFDRAQRACLQNSAIIATPEQLQAAYED-GFHQCDAGW
CORE2   GEVF-YATSPEKFT--FQEAANECRTVGARLATTGQLYLAWQG-GMDMCSAGW
CORE3   GVVFHYRPGSTRYSLTFEEAQQACIRTGAAIASPEQLQAAYEA-GYEQCDAGW
CORE4   GEVF-FATQMEQFT--FQEAQAFCAAQNATLASTGQLYAAWSQ-GLDKCYAGW
LINK1   GVVFPYFPRLGRYNLNFHEARQACLDQDAVIASFDQLYDAWRG-GLDWCNAGW
LINK2   GR-FYYLIHPTK--LTYDEAVQACLNDGAQIAKVGQIFAAWKLLGYDRCDAGW

                                                       *
CD44    IEGHVV-IPRIHPNSICAANNTGVYIL----------TSNTSQYDTYCFNASA
CORE1   LADQTVRYPIHTPREGCYGDKDEFPGVRTYGIR---DTNET--YDVYCFAEEM
CORE2   LADRSVRYPISKARPNCGGNLLGVRTVYLHANQTGYPDPSSR-YDAICYTGED
CORE3   LQDQTVRYPIVSPRTPCVGDKDSSPGVRTYGVR---PSSET--YDVYCYVDKL
CORE4   LADGTLRYPIVNPRPACGGDKPGVRTVYLYPNQTGLPDPLSK-HHAFCFRGVS
LINK1   LSDGSVQYPITKPREPC-GGQNTVPGVRNYGFW---DKDKSR-YDVFCFTSNF
LINK2   LADGSVRYPISRPWRRCSPTEAAVRFV-------GFPDKKHKLYGVYCFRAYN
```

Figure 20. LDLRSF domains

```
                     *                                  *
LDLR-d1       GTAVG--DRCERNEFQCQDG--KCISYKWVCDGSAECQDGSDES
LDLR-d2       ETCLS--VTCKSGDFSCGGRVNRCIPQFWRCDGQVDCDNGSDEQ
LDLR-d3       -GCPP--KTCSQDEFRCHDG--KCISRQFVCDSDRDCLDGSDEA
Comp 9        EQALP--SECSSIEFTCESG--ACIKLRLSCNGDYDCEDGSDED
Hemo. Linker  DELEG--NGCEPRHFQCGGSAMECISDLLTCDGSPDCANGADED
Factor I      ELCC---KACQGKGFHCKSG--VCIPSQYQCNGEVDCITGEDEV
Comp 7        RGCPTE-EGC-GERFRCFSG--QCISKSLVCNGDSDCDEDSADE
Comp 6        LLCKIEEADC-KNKFRCDSG--RCIARKLECNGENDCGDNSDER
```

Figure 21. Ly-6SF domains

```
                *
Human CD59    LQCYNCPN--PTADCKT----------AVNCSSDFDAC---LITK-AGLQ
Mouse Ly-6A   LECYQCYGVPFETSCPS----------ITCPYPDGVC---VTQE-AAVI
Mouse Ly-6C   LQCYECYGVPIETSCPA----------VTCRASDGFC---IAQN-IELI
UPAR-1        LRCMQCKT---NGDCRV----------EECALGQDLCRTTIVRL-WEEG
UPAR-2        LECISCGSS--DMSCERGR----HQSLQCRSPEEQCLD-VVTHWIQEG
UPAR-3        RQCYSCKGN-STHGCSSEE----TFLIDCRGPMNQC---LVAT-GTHE
Squid Sgp2    IKCFVCNSY-HQQDCGDWFDNATHSVHQCEPSQDRCRK-IVQQ-IKLD

                                                                *
V----------YNKCWK---------FEHCNF-NDVTTRLREN--ELTYYCCKKDLCN
VD-SQTRKVKNNLCL----------PICPP-NIESMEILQTKVNVKTSCCQEDLCN
ED-SQRRKLKTRQCL----------SFCPA-GVPIKDPNIR---ERTSCCSEDLCN
EE----LELVEKSCTH----------SEKTNRTLSYRTGLKITSLTEVVCGLDLCN
EEGRPKDDRHLRGCGY--------LPGCPG-SNGFHNNDTF---HFLKCCNTTKCN
PK---NQSYMVRGCATA---------SMCQHAHLGDAFSMNH---IDVSCCTKSGCN
EE-WQVRY--IRQCAEGGEIGAYDGRVCKD-RIGTSGVK-----MTYCHCQTEGCN
```

Figure 20. (opposite) *Low density lipoprotein receptor (LDLR) supe amily domains. Residues identical in four or more sequences are boxed. he asterisks mark the positions of the conserved residues that are marked on t : domain organization figures in the entries in Section II. The sequences of t e following proteins are from the Swissprot database and the database access. n number and residue numbers are given in brackets. LDLR, human low density :poprotein receptor precursor (P01130, d1 20–59, d2 61–102, d3 103–141); Coı ɒ 9, rainbow trout complement C9 (P06682, 72–112); Hemo.Linker, marine wo: ı giant extracellular haemoglobin linker 2 chain (P18208, 61–102); Factoı , human complement factor I precursor (P05156, 253–291); Comp 7, humaı :omplement C7 precursor (P10643, 77–116); Comp 6, human complement C6 ɪ :cursor (P13671, 131–171).*

Figure 21. (opposite) *Ly-6 superfamily domains. Residues identica in three or more sequences are boxed. The asterisks mark the positions of th. :onserved residues that are marked on the domain organization figures in tl entries in Section II. The sequences of the following proteins are from the S' :ssprot database unless otherwise indicated and the database accession ı umber and residue numbers are given in brackets. CD59, human CD59 antig ı (P13987, 26–95); Mouse Ly-6A, (P05533, 27–105); Mouse Ly-6C, (P09568, 2 ·102); UPAR-1, UPAR-2 and UPAR-3, human urokinase plasminogen activator rı eptor (PIR; S12376, 23–99, 115–199 and 214–294 respectively); Squid Sgp2, sç id glycoprotein 2 residues 1–92 [67].*

of the Ly-6 antigens have been found in the rat but not yet in ıther species. In humans the CD59 antigen has been identified as a member of the .y-6 superfamily but seems too different in sequence from the mouse Ly-6 antiı ns to be a Ly-6 homologue. CD59 is a downregulatory control protein for human :omplement and also shows adhesion reactivity with the CD2 antigen [66]. An inv tebrate member of the Ly-6 superfamily has been isolated from squid optic an central nervous tissue [64,67]. All the above molecules consist of a single Ly-6SF dı nain attached to the cell surface by a GPI anchor.

Another member of the Ly-6 superfamily is the urokinase plas: inogen activator receptor. This molecule contains three domains separated by hiı e like sequences and is also attached to the cell surface by a GPI anchor. The ε gnments of this superfamily are shown in Fig. 21 and a tertiary structure for ıis domain type remains to be determined. No Ly-6SF domain has been found in ɔmbination with domains of any other superfamily and this may be because th exon structures known for this superfamily are not suited to exon shuffling (Tabl 1).

The MHC superfamily (Figs 22–24)

The MHC antigens and related molecules are members of the gSF by virtue of their membrane proximal domains being in the IgSF C1 set. Ho ever, their NH_2-terminal segments, including the $\alpha1$ and $\alpha2$ domains of MHC C ıss I heavy chain and the $\alpha1$ and $\beta1$ domains of the Class II α and β chains, ε ow no sequence similarity to IgSF sequences [68] and the Class I domains are k own to form an independent structural unit as shown in Fig. 22 [34]. The Class I c and $\alpha2$ domains show weak sequence similarity to each other and form a simila fold containing a

Figure 20. (opposite) *Low density lipoprotein receptor (LDLR) su rfamily domains. Residues identical in four or more sequences are boxe The asterisks mark the positions of the conserved residues that are marked on he domain organization figures in the entries in Section II. The sequences o he following proteins are from the Swissprot database and the database acce ion number and residue numbers are given in brackets. LDLR, human low densi lipoprotein receptor precursor (P01130, d1 20–59, d2 61–102, d3 103–141); C np 9, rainbow trout complement C9 (P06682, 72–112); Hemo.Linker, marine w m giant extracellular haemoglobin linker 2 chain (P18208, 61–102); Fact I, human complement factor I precursor (P05156, 253–291); Comp 7, hum complement C7 precursor (P10643, 77–116); Comp 6, human complement C6 recursor (P13671, 131–171).*

Figure 21. (opposite) *Ly-6 superfamily domains. Residues identi l in three or more sequences are boxed. The asterisks mark the positions of t conserved residues that are marked on the domain organization figures in e entries in Section II. The sequences of the following proteins are from the vissprot database unless otherwise indicated and the database accessio umber and residue numbers are given in brackets. CD59, human CD59 ant en (P13987, 26–95); Mouse Ly-6A, (P05533, 27–105); Mouse Ly-6C, (P09568, –102); UPAR-1, UPAR-2 and UPAR-3, human urokinase plasminogen activator ceptor (PIR; S12376, 23–99, 115–199 and 214–294 respectively); Squid Sgp2, uid glycoprotein 2 residues 1–92* [67].

of the Ly-6 antigens have been found in the rat but not yet i other species. In humans the CD59 antigen has been identified as a member of t Ly-6 superfamily but seems too different in sequence from the mouse Ly-6 an ens to be a Ly-6 homologue. CD59 is a downregulatory control protein for hum complement and also shows adhesion reactivity with the CD2 antigen [66]. An in rtebrate member of the Ly-6 superfamily has been isolated from squid optic a l central nervous tissue [64,67]. All the above molecules consist of a single Ly-6SF main attached to the cell surface by a GPI anchor.

Another member of the Ly-6 superfamily is the urokinase pla minogen activator receptor. This molecule contains three domains separated by h ge like sequences and is also attached to the cell surface by a GPI anchor. The lignments of this superfamily are shown in Fig. 21 and a tertiary structure fo this domain type remains to be determined. No Ly-6SF domain has been found i combination with domains of any other superfamily and this may be because e exon structures known for this superfamily are not suited to exon shuffling (Ta 4).

The MHC superfamily (Figs 22–24)

The MHC antigens and related molecules are members of th IgSF by virtue of their membrane proximal domains being in the IgSF C1 set. H wever, their NH_2-terminal segments, including the α1 and α2 domains of MHC lass I heavy chain and the α1 and β1 domains of the Class II α and β chains, how no sequence similarity to IgSF sequences [68] and the Class I domains are nown to form an independent structural unit as shown in Fig. 22 [34]. The Class I 1 and α2 domains show weak sequence similarity to each other and form a simil fold containing a

platform of β strands and a single α helix. The two domains together form tl peptide binding groove of the MHC molecule. In the Class II molecules the c domain shows strong sequence similarity to Class I α1 and the Class II β1 doma is most similar to Class I α2 [69]. Thus in the sequence alignments in Figs 23 and : the sequences are shown as an MHC Iα1 set and MHC Iα2 set. These two se

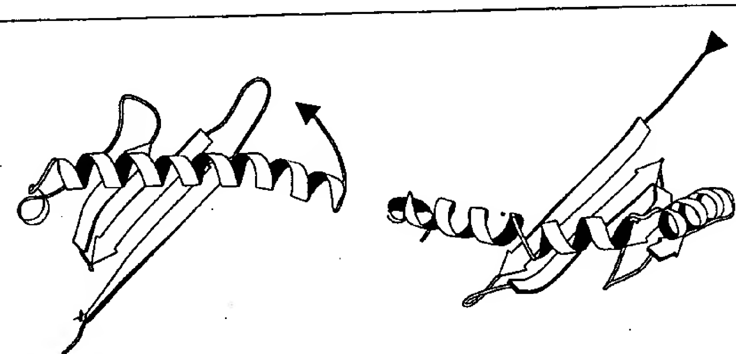


Figure 22. *The folding pattern of MHC superfamily domains. The β strands are shown as broad arrows pointing from the amino to carboxy direction, α helices ⁽ coiled ribbons and the connecting loops as thinner lines. The arrowheads indica where the MHC Class I α1 joins the α2 domain to form the peptide binding groc flanked by the two α helices. The data are from the Brookhaven protein databa:*

Figure 23. (opposite) *MHC Iα1 set superfamily domains. Residues identical in three or more of the sequences are boxed. The positions of the beta strands (β), alpha helices (α) determined for the structure of the human HLA Class I are shown above the sequences. The sequences of the following proteins are from tl Swissprot database and the database accession number and residue numbers a given in brackets. MHC Class I, human HLA Class I A-2 α precursor (P01892, 25–116); CD1A, human CD1A antigen precursor (P06126, 26–112); FcR rat, rat ; Fc receptor precursor (P13599, 25-115); HCMV, human cytomegalovirus glycoprotein H301 precursor (P08560, 19–101); Class II A–B α, mouse MHC Cla II A–B α chain precursor (P14434, 21–103); Class II DQ (3) α, human MHC Cla: II DQ (3) α chain precursor (P01909, 28–109).*

Figure 24. (opposite) *MHC Iα2 set superfamily domains. Residues identical in three or more of the sequences are boxed. The positions of the beta strands (β), alpha helices (α) determined for the structure of the human HLA Class I are shown above the sequences. The sequences of the following proteins are from t Swissprot database and the database accession number and residue numbers a given in brackets. MHC Class I, human HLA Class I A–2 α precursor (P01892, 115–203); CD1A, human CD1A antigen precursor (P06126, 109–199); FcR rat, r gut Fc receptor precursor (P13599, 110–199); HCMV, human cytomegalovirus glycoprotein H301 precursor (P08560, 112–210); Class II A β, mouse H2 Class I A β chain precursor (P14483, 32–122); Class DQ (3) β, human MHC Class II DQ ⟩ β chain precursor (P06126, 109–199).*

together might be ca related to the classi organization, includi with a class II-like o in sequence to MH

Figure 23. MHC class Iα1 set SF domains

together might be called the MHC superfamily. There are numerou sequences related to the classical MHC antigens and these show a Class I tyı structural organization, including the binding of β2-microglobulin, with no exa ples so far with a class II-like organization. The Qa and Tla antigens of mice are ery similar in sequence to MHC Class I antigens whereas the human CD1 an gens show

Figure 23. MHC class Iα1 set SF domains

```
                     β                    β              β
MHC Class 1          G S H S M R Y F F T S - V S R P G R G E P R F I A V G Y V D - D T Q F V R F D S D A A S Q R
CD1A                 S F H V T W I A S F Y - N H S W K Q N L V S G W L S D L Q T H T W D S N S S T I V - - - - -
FcR rat              P R L P L M Y H L A A - V S D L S T G L P S F W A T G W L - G A Q Q Y L T Y N N L - - R Q E
HCMV                 G M H V L R Y G Y T G I F D - - D T S H M T L T V V G I F D G Q H F F T Y H V - - - - - - -
Class II A-B alpha   A D H V G T Y G I S V Y Q S P G D - - - - I G Q Y T F E F D G D E L F Y V D L - - - - - - -
Class II DQ(3)alpha  A D H V A S Y G V N L Y Q S Y G P - - - - S G Q Y T H E F D G D E Q F Y V D L - - - - - - -

   β        α                        α
M E P R A P W I E Q E - G P E Y W D G E T R K V K A H S Q T H R V D L G T L R G Y Y N Q S E A G S
- - F L C P W S R G N F S N E E W K E L E T L F R I R T I R S F E G I R R Y A H E L Q F E Y P F E
A D P C G A W I W E N Q V S W Y W E K E T T D L K S K E Q L F L E A I R T L E N Q I N G T F T L Q
N S S D K A S S R A N G T I - S W M A N V S A A Y P T Y L D G E R A K G - - D L I F N Q T E Q N L
D K K E T V W M L P E F G Q L - A S F D P Q G G L Q N I A V V K H N L G V L T K R S N S T P A T N
G R K E T V W C L P V L R Q - - F R F D P Q F A L T N I A V L K H N L N S L I K R S N S T A A T N
```

Figure 24. MHC class Iα2 set SF domains

```
                     β                    β           β           β
MHC Class 1          G S H T V Q R M Y G C D V G S D W R F L R G Y H Q Y A Y D G K D Y I A L K E D L R S W T A A D
CD1A                 Y P F E I Q V T G G C E L H S G - K V S G S F L Q L A Y Q G S D F V S F Q N N - - S W L P Y P
FcR rat              G T F T L Q G L L G C E L A P D N S S L P T A V - F A L N G E E F M R F N P R T G N W S G E W
HCMV                 S Q S V L T W T H E C N T T E N G S F V A G Y E G F G W D G E T L M E L K D N L T L W T G P N
Class II A beta      R H F V Y Q F M G E C Y F T N G T Q R I R Y V T R Y I Y N R E E Y V R Y D S D V G E H R A V T
Class II DQ(3)beta   E D F V Y Q F K G M C Y F T N G T E R V R L V T R Y I Y N R E E Y A R F D S D V G V Y R A V T

            α              α               α          α
- - - - - V A G N M A K H F C K V L N Q N Q - H E N D I T H N L L S D T C P R F I L G L L D A G K A H L Q
- - - - - - - - - P E T D I V G N L W M K Q P E A A R K E S E F L L T S C P E R L L G H L E R G R Q N L E
Y E I S W L K Q N K T Y I D G K I K N I S E G D T T I Q R N Y L K G N C T Q W S V I Y S G F Q P P V T H
- - - - - - - - - E L G R P D A E Y W N S Q P E I L E R T R A E L D T V C R H N Y E G P E T H T S L R R L
- - - - - - - - - P L G P P A A E Y W N S Q K E V L E R T R A E L D T V C R H N Y Q L E L R T T L Q R R V
```

sequence identity only at the level of about 30%. This level of identity is also see for an Fc receptor of rodent neonatal gut [70] and a Class I-related molecule expresse by cytomegalovirus [71]. Secondary structure predictions have been used to sugges that the 70 kD heat shock proteins (hsp70) may also have a peptide binding groov like the MHC class I antigen; however the hsp70 family of proteins show littl sequence similarity to the other members shown in Figs 23 and 24 [111, 112].

A more detailed discussion of MHC-related sequences can be found in refs. 34 69, 72–74.

Nerve growth factor receptor (NGFR) superfamily (Fig. 25)

Four cysteine-rich repeats were recognized in the extracellular part of the lo affinity nerve growth factor receptor (NGFR) and subsequently related sequenc repeats have been identified in a number of leucocyte cell surface antigen including CD40, MRC OX-40, CD27 and the tumour necrosis factor receptor Figure 25 shows an alignment of some of the repeats. This repeat is unusual in tha most of the NGFRSF molecules contain 3 or 4 repeats. No single NGFRSF repea sequence has been found and the repeat has not been associated with any othe domain types. The gene structure of the NGFR shows that the repeat is not code for by a single exon [75] so it is unlikely that this repeat arose by gene duplication c exons encoding single repeats. A primordial gene with four repeats may hav evolved by unequal crossing-over during recombination and this gene probably gav rise to all known members of the NGFR superfamily by duplication an divergence.

The rhodopsin superfamily (Fig. 26)

The members of this large superfamily of more than 50 proteins are characterize by the presence of seven hydrophobic membrane-spanning sequences and ar reviewed in refs 76, 77. The proteins are oriented with the NH_2-terminus on th extracellular side and the COOH-terminus on the cytoplasmic side of the plasm membrane. Several names have been used to describe this superfamily, such as G protein coupled receptor superfamily, 7TMS (7-transmembrane) and rhodopsi superfamily. We have chosen the term rhodopsin superfamily as this was the firs and best characterized member of this superfamily and does not imply an functional association which might later be shown to be inappropriate. Th sequence conservation is highest in the potential transmembrane segments, wit most diversity in the NH_2- and COOH-termini and the cytoplasmic loop betwee segments 5 and 6. Most members of the rhodopsinSF have been shown to couple t various G-proteins. Experiments using chimeric proteins have shown that th sequences contributing to G-protein attachment are found in transmembran segments 5 and 6 and the cytoplasmic loop between them. A subset of closel related rhodopsinSF members is found on leucocytes and includes the C5aR fMLPR and the IL8R (see the entries in Section II and Fig. 26).

The scavenger receptor (scavengerR) superfamily (Fig. 27)

Three domains with sequence similarities were identified in the extracellula region of the CD5 antigen and later these domains were detected in macrophag scavenger receptors, the complement control protein factor I, the CD6 antigen an the speract receptor protein present in sea urchins [78]. We call this the scavenger superfamily since these molecules are the first of this superfamily with which

Figure 25. *Nerve growth factor receptor (NGFR) superfamily repeats. Residues identical in five or more sequences are boxed. The asterisks mark the positions of the conserved residues that are marked on the domain organization figures in the entries in Section II. The sequences of the following proteins are from the Swissprot database unless otherwise indicated and the database accession number and residue numbers are given in brackets. The sequences are contiguous over the four repeats except for OX-40 which contains a short sequence in place of the third repeat. OX-40, rat MRC OX-40 antigen precursor (P15725, 25–102, 124–164); TNFRI, human tumour necrosis factor receptor precursor I (P19438, 43–196); TNFRII, human tumour necrosis factor receptor precursor II (P20333, 39–201); NGFR, rat nerve growth factor receptor precursor (P07174, 32–190); CD40, human receptor precursor (PIR:S04460; 25–187).*

so seen
pressed
suggest
groove
w little

efs. 34,

ıe low
quence
ıtigens
eptors.
in that
repeat
other
coded
tion of
have
y gave
ı and

erized
id are
n the
lasma
as G-
opsin
first
any
The
with
ween
ıle to
t the
ırane
osely
C5aR,

lular
hage
and
gerR
ch a

Figure 25. *Nerve growth factor receptor (NGFR) superfamily repeats. Residues identical in five or more sequences are boxed. The asterisks mark the positions of the conserved residues that are marked on the domain organization figures in the entries in Section II. The sequences of the following proteins are from the Swissprot database unless otherwise indicated and the database accession number and residue numbers are given in brackets. The sequences are contiguous over the four repeats except for OX-40 which contains a short sequence in place of the third repeat. OX-40, rat MRC OX-40 antigen precursor (P15725, 25–102, 124–164); TNFRI, human tumour necrosis factor receptor precursor I (P19438, 43–196); TNFRII, human tumour necrosis factor receptor precursor II (P20333, 39–201); NGFR, rat nerve growth factor receptor precursor (P07174, 32–190); CD40, human CDw40 antigen precursor (PIR:S04460, 25–187).*

```
              *                                                               *
OX40 (1)     N C V K D T Y P S - - - - - G H K C - - C - - - R E C Q P - - G H G M V S R C D H - T R D T V C H
TNFRI(1)     V C P Q G K Y I H P Q N - - N S I C - - C - - - T K C H K - - G T Y L Y N D C P G P G Q D T D C R
TNFRII(1)    T C R L R E Y Y D Q T - - - A Q M C - - C - - - S K C S P - - G Q H A K V F C T K - T S D T V C V
NGFR (1)     T C S T G L Y T H - - - - - S G E C - - C - - - K A C N L - - G E G V A Q P C G - - A N Q T V C E
CD40 (1)     A C R E K Q Y L I - - - - - N S Q C - - C - - - S L C Q P - - G Q K L V S D C T E - F T E T E C L

OX40 (2)     P C E P G F Y N E A V N Y - - D T C K Q C - - - T Q C N H R S G S E L K Q N C T P - T E D T V C -
TNFRI(2)     E C E S G S F T A S E N H - L R H C L S C - - - S K C R K E M G Q V E I S S C T V - D R D T V C -
TNFRII(2)    S C E D S T Y T Q L W N W - V P E C L S C - - G S R C S S - - D Q V E T Q A C T R - E Q N R I C -
NGFR (2)     P C L D N V T F S D V V S A T E P C K P C - - - T E C L G - - L Q S M S A P C V E - A D D A V C -
CD40 (2)     P C G E S E F L D T W N R E - T H C H Q H - - - K Y C D P N L G L R V Q Q K G T S - E T D T I C -

TNFRI(3)     G C R K N Q Y R H Y W S E N L F Q C F N C - - - S L C L N - - G T V H L S - C Q E - K Q N T V C T
TNFRII(3)    T C R P G W Y C A L S K Q - - E G C R L C A P L R K C R P - - G F G V A R P G T E - T S D V V C K
NGFR (3)     R C A Y G Y Y Q D E E T - - - G H C E A C - - - S V C E V - - G S G L V F S C Q D - K Q N T V C E
CD40 (3)     T C E E G W H C T S E - - - - - A C E S C V L H R S C S P - - G F G V K Q I A T G - V S D T I C E

OX40 (4)     P C P P G H F S P G S N Q - - - A C K P W - - - T N C T L - S G K Q I R H P A S N - S L D T V C E
TNFRI(4)     - C H A G F F L R E N - - - - - E C V S C - - - S N C K K - - - - - - S L E C T K - - - - - L C L
TNFRII(4)    P C A P G T F S N T T S S - T D I C R P H - - - Q I C N V - - - - - V A I P G N A - S M D A V C T

CD40 (4)     P C P V G F F S N V S S A - F E K C H P W - - - T S C E T - K D L V V Q Q A G T N - K T D V V C G
```

Figure 26. Rhodopsin superfamily

```
                                                                                       I
IL8R        MSNITDPQMWDFDDLN----FTGMPPADEDYSPCMLETETLNKYVVIIAYALVFLLSLLGNSLVMLVILYSRVGR
C5aR        ----MNSFNYTTPDYGHYDDKDTLDLNTPVDKT--SNTLRVPDILALVIFAVVFLVGVLGNALVVWVTAF-EAKR
fMLPR       -----------------METNSSLPTNISGGTPAVSAGYLFLDIITYLVFAVTFVLGVLGNGLVIWVAGF-RMTH
Rhodopsin   -----MNGTEGPNFYVPFSNATGVVRSPFEYPQYYLAEPWQFSMLAAYMFLLI-VLGFPINFLTLYVTVQHKKLR
NeurokininR -MGTCDIVTEANISSGPESNTTGIT--------AFSMPSWQLALWAPAYLALV-LVAVTGNAIVIWIILAHRRMR
DopaR       -----------------------MRTLNTSAMDGTGLVVERDFSVRILTACFLSLLILSTLLGNTLVCAAVIRFRHLR

            II                                          III
S-VTDVYLLNLALADILFALTLPIWAASKVNG--WIFGTFLCKVVSLLKEVNFYSGILLLACISVDRYLAIVHATRTLTQKRH-LV
T-INAIWFLNLAVADFLSCLALPILFTSIVQHHHWPFGGAACSILPSLILLNMYASILLLATISADRFLLVFKPIWCQNFRGAGLA
T-VTTISYLNLAVADFCFTSTLPFFMVRKAMGGHWPFGWFLCKFLFTIVDINLFGSVFLIALIALDRCVCVLHPVWTQNHRTVSLA
T-PLNYILLNLAVADLFMVLGGFTSTLYTSLHGYFVFGPTGCNLEGFFATLGGEIALWSLVVLAIERYVVVCKPMSNFRFGENH-A
T-VTNYFIVNLALADLCMAAFNAAFNFVYASHNIWYFGRAFCYFQNLFPITAMFVSIYSMTAIAADRYMAIVHPFQP-RLSAPS-T
SKVTNFFVISLAVSDLLVAVLVMPWKAVAEIAGFWPFGS-FCNIWVAFDIMCSTASILNLCVISVDRYWAISSPFRYERKMTPKAA

      IV                                                                  V
KFVCLGCWGLSMNLSLPFFLFRQAYHPNNSSPV-CYEVLGNDTAKWR---------MVLRILPHTFGFIVPLFVMLFCYGFTLRT
WIACAVAWGLALLLTIPSFLYRVVREEYFPPKVLCGVDYSH--DKRRER---------AVAIVRLVLGFLWPLLTLTICYTFILLR
KKVIIGPWVMALLLTLPVII-RVTTVPGKTGTVACTFNFSPWTNDPKERINVAVAMLTVRGIIRFIIGFSAPMSIVAVSYGLIATK
IMGVAFTWVMALACAAPPLAGWSRYIPEGLQCS-CGDYI---------YTLKPEVNNESFVIYMFVVHFTIPMIIIFFCYGQLVFT
KAVIAGIWLVALA-LASPQCFYSTVTMDQGATK-CVAWV---------PEDSGGKTLLLYHLVVIALIYFLPLAVMFVAYSVIGLT
FILISVAWTLSVLISFIPVQLSWHKAKPTSPSDGNATSLAETID------NCDSSLSRTYAISSSVISFYIPVAIMIVTYTRIYRI

                                                                VI
LFK-------------------------------AHMGQKHRAMRVIFAVVLIFLLCWLPYNLVLLADTLMRTQVIQET
TWS-------------------------------RRATRSTKTLKVVVAVVASFFIFWLPYQVTGIM--MSFLEPSSPT
IHK-------------------------------QGLIKSSRPLRVLSFVAAAFFLCWSPYQVVALIATVRIRELLQGM
V-----------------------KEAAAQQQESATTQ--KAEKEVTRMVIIMVIAFLICWVPYASVAFYIFTHQG------
LWR-----------------------RAVPGHQAHGANLRHLQAKKKFVKTMVLVVLTFAICWLPYHLYFILGSFQEDIYCHK-
AQKQIRRIAALERAAVHAKNCQTTTGNGKPVECSQPESSFKMSFKRETKVLKTLSVIMGVFVCCWLPFFILNCILPFCGSGETQPF

                 VII
CERRNNIGRALDATEILGFLHSCLNPIIYAFIGQNFRHGFLKILA--MHGLVSKEFLARHRVT--SYTSSSVNVSSNL-------
F---LLLNKLDSLCVSFAYINCCINPIIYVVAGQGFQGRLRKSLPSLLRNVLTEESVVRESKSFTRSTVDTMAQKTQAV-------
Y---KEIGIAVDVTSALAFFNSCLNPMLYVFMGQDFRERLIHALPASLERALTEDS-TQTSDTATNSTLPSAEVALQAK-------
SNFGPIF---MTIPAFFAKSAAIYNPVIYIMMNKQFRNCMLTTI-CCGK-NPLGDDEASATVSKTETSQVAPA-----------
-FIQQVY----LALFWLAMSSTMYNPIIYCCLNHRFRSGFRLAFRCCPWVTPTKEDKLELTPTTSLSTRVNRCHTKETL   41 amino acids
CIDSNTF----DVFVWFGWANSSLNPIIYAF-NADFRKAFSTLLGCYRLCPATNNAIETVSINNNGAAMFSSHHEPRGS   66 amino acids
```

Figure 26. (above) *Rhodopsin superfamily. Residues identical in four or more sequences are boxed. The bars over the sequences indicate the transmembrane regions. The sequences of the following proteins are from the Swissprot database unless otherwise indicated and the database accession number. IL8R, human high affinity IL8 receptor [109]; C5aR, human C5a anaphylatoxin chemotactic receptor (P21730); fMLPR, human fMet-Leu-Phe receptor (P21462); Rhodopsin, human rhodopsin (P08100); NeurokininR, human neurokinin A receptor (P21452); DopaR, human D(1) dopamine receptor (P21728).*

Figure 27. (below) *Scavenger receptor superfamily domain. Residues identical in four or more sequences are boxed. The asterisks mark the positions of the conserved residues that are marked on the domain organization figures in the entries in Section II. The sequences of the following proteins are from the Swissprot database unless otherwise indicated and the database accession*

```
. . . LNPIIYAFIGQNFRHGFLKILA--MHGLVSKEFLARHRVT---SYTSSSVNVSSNL-------
F---LLLNKLDSLCVSFAYINCCINPIIYVVAGQGFQGRLRKSLPSLLRNVLTEESVVRESKSFTRSTVDTMAQKTQAV-------
Y---KEIGIAVDVTSALAFFNSCLNPMLYVFMGQDFRERLIHALPASLERALTEDS-TQTSDTATNSTLPSAEVALQAK-------
SNFGPIF---MTIPAFFAKSAAIYNPVIYIMMNKQFRNCMLTTI-CCGK-NPLGDDEASATVSKTETSQVAPA-----------
-FIQQVY----LALFWLAMSSTMYNPIIYCCLNHRFRSGFRLAFRCCPWVTPTKEDKLELTPTTSLSTRVNRCHTKETL  41 amino acids
CIDSNTF----DVFVWFGWANSSLNPIIYAF-NADFRKAFSTLLGCYRLCPATNNAIETVSINNNGAAMFSSHHEPRGS  66 amino acids
```

Figure 26. (above) *Rhodopsin superfamily. Residues identical in four or more sequences are boxed. The bars over the sequences indicate the transmembrane regions. The sequences of the following proteins are from the Swissprot database unless otherwise indicated and the database accession number. IL8R, human high affinity IL8 receptor* [109]*; C5aR, human C5a anaphylatoxin chemotactic receptor (P21730); fMLPR, human fMet-Leu-Phe receptor (P21462); Rhodopsin, human rhodopsin (P08100); NeurokininR, human neurokinin A receptor (P21452); DopaR, human D(1) dopamine receptor (P21728).*

Figure 27. (below) *Scavenger receptor superfamily domain. Residues identical in four or more sequences are boxed. The asterisks mark the positions of the conserved residues that are marked on the domain organization figures in the entries in Section II. The sequences of the following proteins are from the Swissprot database unless otherwise indicated and the database accession number and residue numbers are given in brackets. SREC, sea urchin egg peptide speract receptor precursor (P16264, d1 38–145, d2 148–258, d3 259–367, d4 377–486); CD5, human CD5 antigen (P06127, d1 30–134, d2 156–269, d3 271–369); CFAI, human complement factor I precursor (P05156, 109–216); SCAV, human scavenger receptor type I (P21757, 345–451). Alignments are from 20 residues from the conserved glycine to the COOH-terminus of the scavenger receptor.*

```
                                 *                                                                   
SREC d4   PKHTRVRIV-GMGQQGRVEVSLGNGWGR--VCDPDW---------SDHEAKTVCYHAGYKWGA
CD5 d1    DPDFQARLTRSNSKCQGQLEVYLKDGWHM--VCSQSWGRSSKQWEDPSQASKVCQRLNCG---
CFAI      EGKFSVSLKHGNTDSEGIVEVKLVDQDKTMFICKSSW---------SMREANVACLDLGFQQGA
CD5 d2    PPRLQLVAQSGGQHCAGVVEFYSGSLGGT--ISYEAQD------KTQDLENFLCNNLQCOSFL
CD5 d3    QPKVQSRLVGGSSICEGTVEVRQGAQWAA--LCDSSSARS-------SLRWEEVCREQQCG---
SREC d1   AVEGNIRLIHGRTENEGSVEIYHATRWGG--VCDWWW---------HMENANVTCKQLGFP-GA
SREC d2   EPQGSLRMILGDVPNEGTLETFWDGAWGS--VCHTDF---------GTPDGNVACRQMGYS-RG
SREC d3   NVEGDIRLMDGSGPHEGRVEIWHDDAWGT--ICDDGW---------DWADANVVCRQAGYR-GA
SCAV      TPFTKVRLVGGSGPHEGRVEILHSGQWGT--ICDDRW---------EVRVGQVVCRSLGYP-GV

                                                                       *
. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
DTQRRFKLS---DLSINSTEC---LHVHCRGLET-SLAEC---TFTKRRTMGY--QDFADVVCYT
KHLPETEAGRAQDPGEPREHQPLPIQWKIQNSSCTSLEHCFRKIKPQKSGR------VLALLCSG
VNSYR-------VLDAGDPTS---RGLFC---PHQKLSQCH-ELWERNSYCK-----KVFVTCQD
RQFYRR-----AYFGAHVTTFWVY-KMNCLGNET-RLEDCYHRPYGRPWLC--NAQWAAGVECLP
VKSIKT----DGHFGFSTGPI-ILDAVDCEGTEA-HITECNMPVTPYQHACPYTHNWDVGVVCKP
VKASGFK---GEDFGFTWAPI-HTSFVMCTGVED-RLIDCILRD-GWTHSC--YHVEDASVVCAT
QAVHKA-----AHFGQGTGPI-WLNEVFCFGRES-SIEECKIRQWG-TRAC--SHSEDAGVTCTL
```

clear functional activity has been associated. However, some ligands will bind to both scavenger receptors I and II but the latter lacks this type of domain so the functional involvement of this domain remains to be resolved [79]. Initially it was argued that the CD5 antigen domains were related to IgSF domains [80]. However, this contention was not supported by ALIGN analysis as described above. Subsequently it was suggested that the CD5 domains were related in sequence to the domains of the PapD bacterial protein [11] but again this was not supported by a detailed analysis including CD5 domains plus numerous other scavengerRSF domains. It is now clear that there is a separate superfamily of proteins containing scavengerRSF domains and alignments for this superfamily are shown in Fig. 27. No tertiary structure data are available yet for these domains.

Signal transduction sequence motifs (Fig. 28)

The signal transduction sequence motif shown in alignments in Fig. 28 is present in the cytoplasmic regions of several membrane proteins present in the antigen receptor complexes on B cells, T cells, and the IgE receptor on mast cells [81]. This motif is also found in the CD5 antigen cytoplasmic domain (Beyers, Spruyt and Williams, unpublished). The CD3 ζ chain is unusual in that it has three motifs whereas the others have only one. One common feature of these molecules is that they are components of membrane complexes which, when cross-linked, give signals that lead to cell activation. This results in cell proliferation in the case of the antigen receptors and to degranulation of mast cells. Cross-linking of CD3ε by

```
Human CD3 gamma                DKQTLLPNDQLYQPLKDREDDQ-YSHLQQNQLRRN
Human CD3 delta                DTQALLRNDQVYQPLRDRDDAQ-YSHLGGNWARNK
Mouse CD3 epsilon              KERPPPVPNPDYEPIRKGQRDL-YSGL--------
Human CD3 zeta (1)             PPAYQQGQNQLYNELNLGRREE-YDVLDKRRGRDP
Human CD3 zeta (2)             KPRRKNPQEGLYNELQKDKMAEAYSEIGMKG----
Human CD3 zeta (3)             ERRRGKGHDGLYQGLSTATKDT-YDALHMQALPPR
Mouse MB1                      DMPDDYEDENLYEGLNLDDCSM-YEDISRGLQGTY
Mouse B29                      DGKAGMEEDHTYEGLNIDQTAT-YEDIVTLRTGEV
Rat Fc epsilon R beta chain    FERSKVPDDRLYEELHVYSPI--YSALEDTREASA
Rat Fc epsilon R gamma chain   DIASREKSDAVYTGLNTRNQET-YETLKHEKPPQ-
Human CD5                      ENPTASHVDNEYSQPPRNSRLSAYPALEGVLHRS-
```

Figure 28. *Signal transduction motifs. Residues identical in five or more sequences are boxed. The sequences of the following proteins are from the Swissprot database and the database accession number and residue numbers are given in brackets. CD3 gamma, human CD γ chain precursor (P09693, 149–181); CD3 delta, human CD3 δ chain precursor (P04234, 138–171); CD3 epsilon, mouse CD3 ε chain precursor (P22646, 159–184); CD3 zeta, human CD3 ζ chain precursor (P20963, 61–94, 99–130, 131–163); MB1, mouse MB-1 protein precursor (P11911, 171–204); B29, mouse B cell glycoprotein B29 precursor (P15530, 184–217); Fc epsilon R beta, rat Ig ε receptor β subunit (P13386, 207–239); Fc epsilon R gamma, rat Ig ε receptor γ subunit precursor (P20411, 54–86); CD5, human CD5 antigen precursor (P06127, 442–475).*

PC1 d1	K S C - K G R C F E - - R T F S N C R C D A A C V S L G N C C L D	Q E T C V E P T H
PP11	T S C - Q G R C Y E A F D K H H Q C H C N A R C Q E F G N C C K D	E S L C S D H E V
Vitronectin	E S C - K G R C T E G F N V D K K C Q C D E L C S Y Y Q S C C T D	T A E C K P Q V T
PC1 d2	W T C N K F R C G E K R L S R F V C S C A D D C K T H N D C C I N	S S V C Q D K K S

Figure 29. *Somatomedin B superfamily domains. Residues identical : three or more sequences are boxed. The asterisks mark the positions of the cc served residues that are marked on the domain organization figures in the e tries in Section II. The sequences of the following proteins are from the Swis. rot database and the database accession number and residue numbers c given in brackets. PC1, mouse plasma cell antigen PC1 (P06802; d1, 54–93; d. 95–137); PP11, human placental protein precursor (P21128, 47–88); Vitronecti human vitronectin precursor (P04004; 22–63).*

immobilized mAbs leads to T cell proliferation. The use of chimeri proteins has shown that the cross-linking of the cytoplasmic domains of the T ll receptor ζ chain or the CD3ε, gives a TcR-like signal in cells lacking surface ex ession of the TcR, implying that this motif is involved in coupling the TcR t intracellular transduction mechanisms [82,83].

Somatomedin B superfamily (Fig. 29)

Somatomedin B is a serum peptide derived from vitronectin (alsc called serum spreading factor) by proteolysis. The plasma cell surface antigen PC- was noted to contain two somatomedin BSF repeats [41]. This glycoprotein l s nucleotide pyrophosphatase/alkaline phosphodiesterase activity [84] but this is a ociated with a different region of the molecule than that containing the som tomedin BSF repeat. The domain has not been found on other cell surface molecu s although it is present in placental protein 11 (PP11) [85].

Transmembrane 4 pass (TM4) superfamily and the relationship betw n FcεRIβ Chain and CD20 (Figs 30 and 31)

The "TM4 superfamily" is a term that we suggest for a new group o proteins with clear sequence similarities that are thought to traverse the lipid bil er four times with both the NH_2- and COOH-termini on the cytoplasmic face of e membrane. This superfamily includes several leucocyte antigens such as CD9 CD37, CD53, CD63 and TAPA-1. Alignments for the TM4 superfamily are shown n Fig. 30. The genomic sequence of the TAPA-1 antigen shows that the sequen : is coded by eight exons which do not indicate any simple correlation witl the proposed transmembrane sequences [86]. The intron/exon boundaries of mc se CD53 are largely compatible with those of TAPA-1 [87] in support of the argur nt that these molecules had a common ancestor in evolution. The majority of th differences in sequence between TM4 superfamily molecules reside in the ex cellular loop between TM sequences 3 and 4 where there are considerable lifferences in sequence length. This loop of sequence is known to be extracell lar because it includes the N-linked glycosylation sites and the MRC OX-44 epit e that can be labelled at the cell surface maps to an Ile/Thr interchange in tl region [87]. In addition, surface labelling studies on TAPA-1 support an extracell r localization

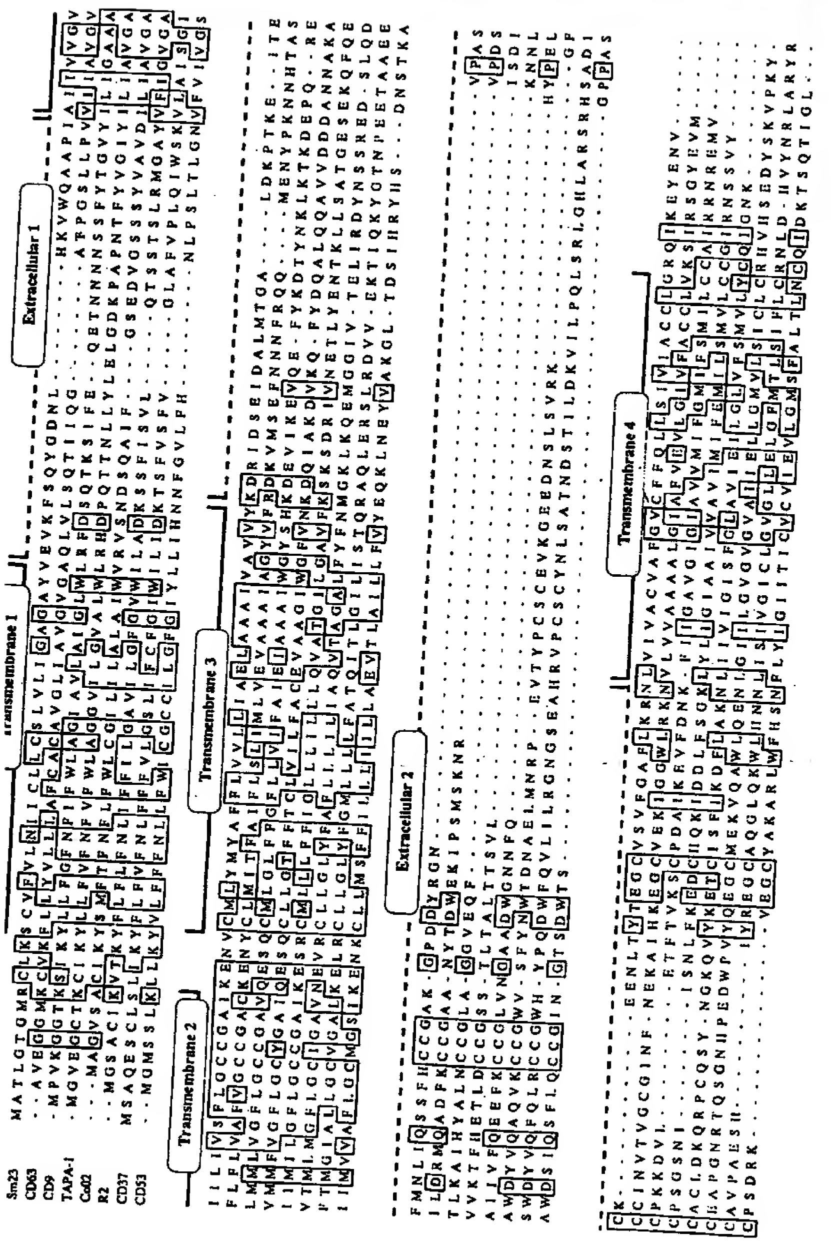


Figure 30. (opposite) *Transmembrane 4*
in four or more sequences are boxed. T
from the Swissprot database or the refe
number is given in brackets (the compl
Schistosoma mansoni *protein Sm23 (P*
antigen ME491 (P08962); CD9, human
TAPA-1 antigen (P18582); Co029, hum
(P19075); R2, human R2 antigen [110]*; C*
human CD53 antigen (P19397).

Figure 31. (below) *Alignment of CD20 a*
between the two sequences are boxed.
transmembrane sequences are indicate
The similarities are mostly within or a
fall off towards the COOH-terminus. T
and an ALIGN score of 6.1 SD was obta
Swissprot database accession numbers
chain (P20490).

```
CD20     MTT---PRNSVNGTFPA
FcER1B   MDTENKSRADLALPNPQ

SLVGPT--QSFFMREESKTLGAVQ
SPPPQQTWQSFLKKELEFLGVTQ

APICVT---VWYPLWGGIMYIIS
DDEVLLLYRAGYPFWGAVLFVLS

SLSLFAAISGMILSIMDILNIKI
IVSSIAA--GLGIAIL-ILNLS-

CEPANPSEKNSPSTQYCYSIQSL
C-------KDITEDDGCF-VTS-

ENEWKRTCSRPKSNCIVLLSAEE
LIIY-------------------

NEEDIEII-PIQEEEEEETETNF
LYEELHVYSPIYSALEDTREAF-
```

for this loop [88]. There is no known fu
although antibodies against members
proliferation of leucocytes [89].

There are other leucocyte proteins t
membrane four times, including CD20
show sequence similarity to the TM4 su
chain show clear sequence similaritie
transmembrane regions of these molecul

Figure 30. (opposite) *Transmembrane 4 pass (TM4) superfamily. Residues identical in four or more sequences are boxed. The sequences of the following proteins are from the Swissprot database or the reference given and the database accession number is given in brackets (the complete sequences are shown). Sm23,* Schistosoma mansoni *protein Sm23 (P19331); CD63, human melanoma associated antigen ME491 (P08962); CD9, human CD9 antigen (P21926); TAPA-1, human TAPA-1 antigen (P18582); Co029, human tumour associated antigen Co-029 (P19075); R2, human R2 antigen* [110]*; CD37, human CD37 antigen (P11049); CD53, human CD53 antigen (P19397).*

Figure 31. (below) *Alignment of CD20 and FcεR1 β chain. Residues identical between the two sequences are boxed. The possible positions of the four transmembrane sequences are indicated with a bar above or below the sequences. The similarities are mostly within or around the transmembrane sequences and fall off towards the COOH-terminus. There are many conservative substitutions and an ALIGN score of 6.1 SD was obtained for the full sequences shown. The Swissprot database accession numbers are; human CD20 (P22836), mouse FcεRI β chain (P20490).*

```
CD20     MTT---PRNSVNGTFPAEPMKGP-IAM-QSGP--KPLFRRMS
FcER1B   MDTENKSRADLALPNPQESPSAPDIELLEASPPAKALPEKPA

SLVQPT--QSFFMRESKTLGAVQIMNGLFHIALGGLL--MIPAGIY
SPPPQQTWQSFLKKELEFLGVTQVLVGLICLCFGTVVCSTLQTSDF

APICVT---VWYPLWGGIMYIISGSLLAATEKNSRKCLVKGKMIMN
DDEVLLLYRAGYPFWGAVLFVLSGFLSIMSERKNTLYLVRGSLGAN

SLSLFAAISGMILSIMDILNIKISHFLKMESLNFIRAHTPYINIYN
IVSSIAA--GLGIAIL-ILNLS----------------NNSAYMN-Y-

CEPANPSEKNSPSTQYCYSIQSLFLGILSVMLIFAFFQELVIAGIV
C-------KDITEDDGCF-VTS-FITELVLMLLFLTILAFCSAFVL

ENEWKRTCSRPKSNCIVLLSAEEKKEQTIEIKEEVVGLTETSSQPK
LIIY-------------------RIGQEFE-RSKV-------PDDR

NEEDIEII-PIQEEEEEETETNFPEPPQDCQESSPIENDSSP
LYEELHVYSPIYSALEDTREAF----------SAPVVS----
```

for this loop [88]. There is no known function for any of this family of proteins although antibodies against members of this family do have effects on the proliferation of leucocytes [89].

There are other leucocyte proteins that are predicted to traverse the plasma membrane four times, including CD20 and the FcεRIβ chain, but these do not show sequence similarity to the TM4 superfamily. However, CD20 and the FcεRIβ chain show clear sequence similarities to each other [90] in three of the four transmembrane regions of these molecules as shown in Fig. 31, and their genes are

very closely linked on mouse chromosome 19 [90]. Thus these two sequences should be considered as founder members of a new superfamily and perhaps for now this could be referred to as the FcεRI β superfamily. There are data to show that CD20 is a Ca^{2+} channel [91] and it would be interesting to know if this is also the case for FcεRI β chain and members of the TM4SF too.

The tyrosine kinase superfamily (Fig. 32)
Tyrosine kinase domains are found in the cytoplasm and two groups can be distinguished: receptor tyrosine kinases which are transmembrane proteins, and nonreceptor tyrosine kinases which are located in the cytoplasm. The non-receptor group of kinases includes members of the src family, all of which are anchored to the inner leaflet of the plasma membrane with a myristate moiety. On activation they phosphorylate Tyr residues on their own cytoplasmic domains or on other proteins in the cytoplasm and this is believed to be one of the key early events in signal transduction pathways after ligand recognition. In leucocytes the best studied example is p56^{lck} which associates with the cytoplasmic domains of CD4 and CD8 and regulates signal transduction by these molecules [92]. Other examples are *fyn* which associates with the T cell receptor complex [93], and *lyn*, *fyn*, and *blk*, which couple to the membrane Ig complex of B cells [94,95].

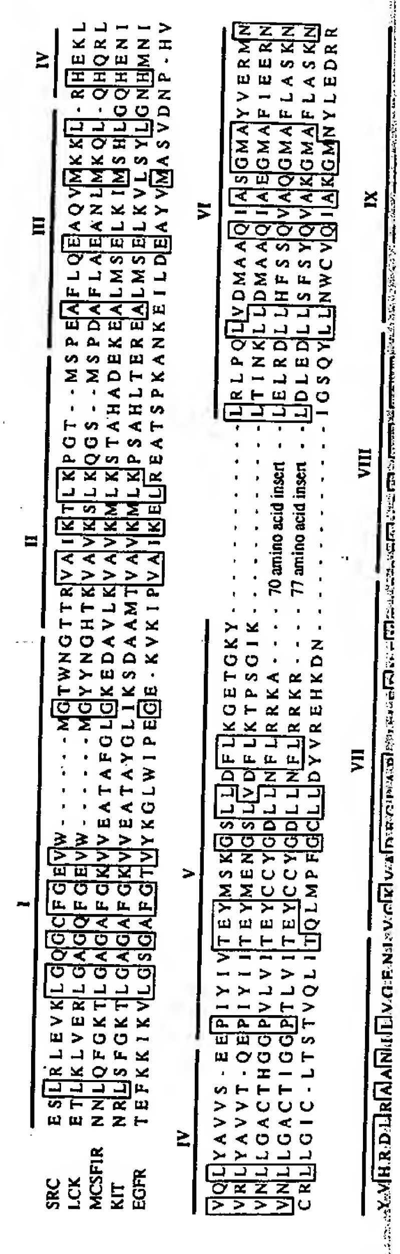
Receptor tyrosine kinases are expressed on a wide variety of cells and examples nclude the PDGF receptor and EGF receptor. When these receptors bind their natural ligands they oligomerize and the cytoplasmic tyrosine kinase domains become activated and autophosphorylated. This leads to the phosphorylation and ictivation of various intracellular substrates including phospholipase Cγ , phosphatidylinositol 3-kinase and the c-raf serine kinase. These effector molecules oncomitantly associate with the activated receptor kinases [96,97].

Tyrosine kinase domains consist of about 260–360 amino acids. The difference in ize is due to insertion of a "kinase insert domain" of about 70–100 amino acids in ertain receptor kinases, including the platelet derived growth factor receptor, M-CSFR, and c-kit kinases. These insert regions appear to regulate the interaction of ne kinase with certain cellular substrates/effector molecules [96,97]. The tyrosine inase domain of a particular molecule is particularly well-conserved across species nd the identities between molecules within the superfamily are about 40%, as lustrated in Fig. 32. This is much higher than for many of the superfamilies with omains that are found at the cell surface.

Kinase domains are not conserved uniformly, but consist of 11 highly conserved ibdomains (I–XI) separated by regions of lower conservation [98]. Subdomain I ntains the Gly-X-Gly-X-X-Gly consensus which forms part of the binding site for TP. Subdomain II contains an invariant lysine, which appears to be directly volved in the phosphotransfer reaction. Subdomain VIII contains a Pro-Ile-/Val-s/Arg-Trp-Thr/Met-Ala-Pro-Glu consensus which is characteristic of the tyrosine nases. In the serine/threonine kinases the consensus is Gly-Thr/Ser-X-X-Tyr/Phe-Ala-Pro-Glu.

Tyrosine kinases have been the subject of extensive study in nonlymphoid cells d the subject has been reviewed in depth [96,98,99].

e phospho-tyrosine phosphatase (PTPase) superfamily (Fig. 33)
e PTPase superfamily of integral membrane proteins was discovered when two oplasmic homology units of the CD45 antigen [100] were matched with the



```
                          I                          II              III         IV
SRC     ESLRLEVKLGQGCFGEVW------MGTWNGTTRVAIKTLKPGT--MSPEAFLQEAQVMKKL-RHEKL
LCK     ETLKLVERLGAGQFGEVW------MGYYNGHTKVAVKSLKQGS--MSPDAFLAEANLMKQL-QHQRL
MCSFIR  NNLQFGKTLGAGAFGKVVEATAFGLGKEDAVLKVAVKMLKSTAHADEKEALMSELKIMSHLGQHENI
KIT     NRLSFGKTLGAGAFGKVVEATAYGLIKSDAAMTVAVKMLKPSAHLTEREALMSELKVLSYLGNHMNI
EGFR    TEFKKIKVLGSGAFGTVYKGLWIPEGE-KVKI-PVAIKELREATSPKANKEILDEAYVMASVDNP-HV

   IV            V                                                     VI
VQLYAVVS-EEPIYIVTEYMSKGSLLDFLKGETGKY------------LRLPQLVDMAAQIASGMAYVERMN
VRLYAVVT-QEPIYIITEYMENGSLVDFLKTPSGIK------------LTINKLLDMAAQIAEGMAFIEERN
VNLLGACTHGGPVLVITEYCCYGDLLNFLRRKA---- 70 amino acid insert --LELRDLLHFSSQVAQGMAFLASKN
VNLLGACTIGGPTLVITEYCCYGDLLNFLRRKR---- 77 amino acid insert --LDLEDLLSFSYQVAKGMAFLASKN
CRLLGIC-LTSTVQLITQLMPFGCLLDYVREHKDN------------IGSQYLLNWCVQIAKGMNYLEDRR

                             VII                 VIII              IX
YVHRDLRAANILVGENLVCKVADFGLAR-LIEDNEYTARQGAKFPIKWTAPEAALYGRFTIKSDVWSFGILL
YIHRDLRAANILVSDTLSCKIADFGLAR-LIEDNEYTAREGAKFPIKWTAPEAINYGTFTIKSDVWSFGILL
CIHRDVAARNVLLTNGHVAKIGDFGLARDIMNDSNYIVKGNARLPVKWMAPESIFDCVYTVQSDVWSYGILL
CIHRDLAARNILLTHGRITKICDFGLARDIKNDSNYVVKGNARLPVKWMAPESIFNCVYTFESDVWSYGIFL
LVHRDLAARNVLVKTPQHVKITDFGLAKLLGAEEKEYHAEGGKVPIKWMALESILHRIYTHQSDVWSYGVTV

                    X                                  XI
TELTTKGRVPYPGM-VNREVLDQVERGYRMPCPPECPESLHDLMCQCWRKEPEERPTFEYLQAFLEDYFTST
TEIVTHGRIPYPGM-TNPEVIQNLERGYRMVRPDNCPEELYQLMRLCWKERPEDRPTFDYLRSVLEDFFTAT
WEIFSLGLNPYPGILVNSKFYKLVKDGYQMAQPAFAPKNIYSIMQACWALEPTHRPTFQQICSFLQEQAQED
WELFSLGSSPYPGMPVDSKFYKMIKEGFRMLSPEHAPAEMYDIMKTCWDADPLKRPTFKQIVQLIEKQISES
WELMTFGSKPYDGIPA-SEISSILEKGERLPQPPICTIDVYMIMVKCWMIDADSRPKFRELIIEFSKMARDP
```

Figure 32. *Tyrosine kinase superfamily. Residues identical in four or more sequences are boxed. The bars above the sequences represent the subdomains defined in this superfamily and the numbering is as in ref. 98. The sequences of the following proteins are from the Swissprot database and the database accession number and residue numbers are given in brackets. SRC, human src proto-oncogene tyrosine kinase (P12931, 268–526); LCK, human T cell specific tyrosine kinase (P06239, 243–501); MCSFR, human macrophage colony stimulating factor receptor precursor (P07333, 580–917); KIT, human kit proto-oncogene precursor (P10721, 587–931); EGFR, human EGF receptor precursor (P00533, 710–975).*

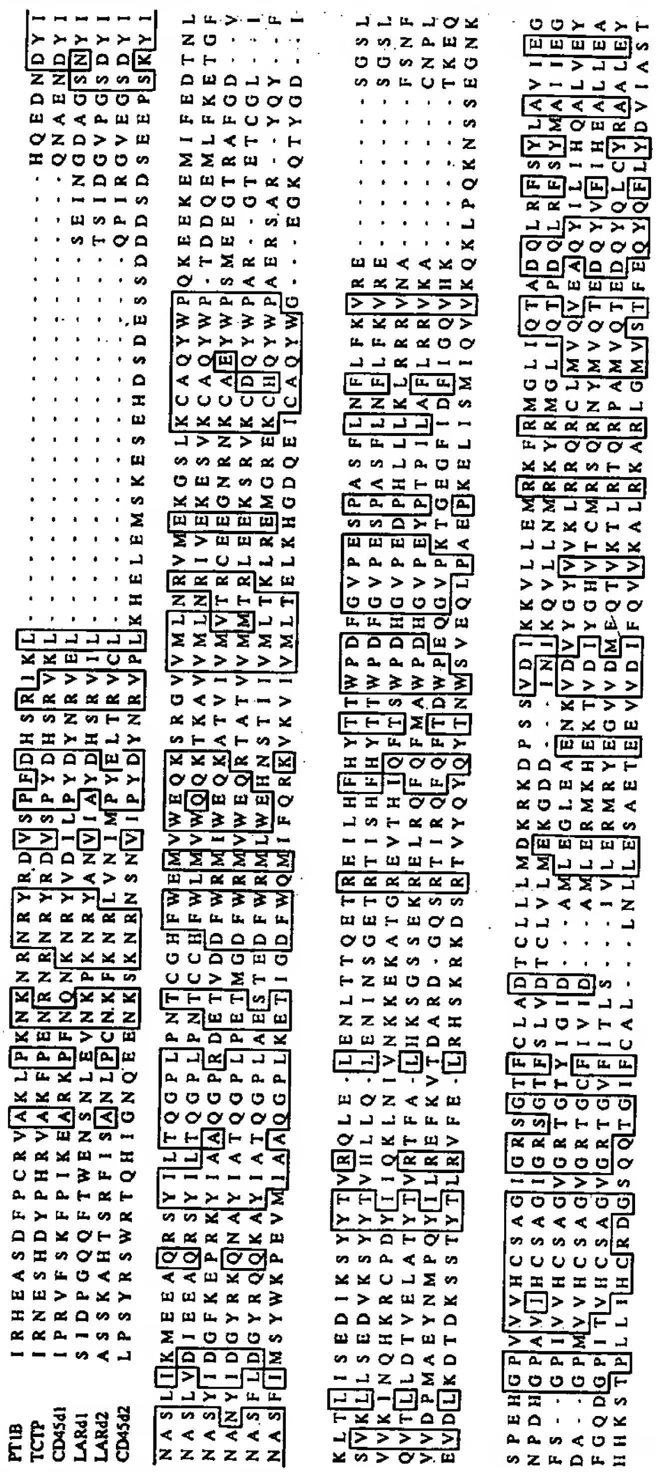


Figure 33. *Phospho-tyrosine phosphatase (PTPase) superfamily domains. Residues identical in four or more sequences are boxed. The sequences of the following proteins are from the Swissprot database and the database accession number and residue numbers are given in brackets. PT1B, human protein phosphotyrosine phosphatase PT1-B (P18031, 23–277); TCTP, human T cell protein phosphotyrosine phosphatase (P17706, 25–275); CD45, human CD45 antigen (P08575, d1 658–910, d2 949–1226); LAR, human LAR protein (P10586, d1 1348–1597, d2 1637–1888).*

sequence of a placental cytop
sequently PTPase activity has b
domain of CD45 but as yet not
other sequences have been ident
hybridization with cDNA probe
surface protein with three IgSF
extracellular side, although this
Sequences with many similariti
Drosophila [105]. Some examples
domain in CD45 is unusual in c
in that it contains an insertion c
and Ser residues. The Ser residu
an extremely negatively charged

The only complete genomic
CD45 [106]. This shows that the
exons for each domain. Howe
sequence similarities do not co
of intron/exon boundaries. The

References

1 Dayhoff, M.O. et al. (1983) M
2 Williams, A.F. and Barclay, A
3 Baron, M. et al. (1991) Trend
4 Pearson, W.R. and Lipman, D
5 Patthy, L. (1990) Cell 61, 13–
6 Williams, A.F. (1987) Immun
7 Williams, A.F. et al. (1989) C
8 Ryu, S.E. et al. (1990) Nature
9 Wang, J. et al. (1990) Nature
10 Driscoll, P.C. et al. (1991) Na
11 Holmgren, A. and Branden, C
12 Amzel, L.M. and Poljak, R.J.
13 de Vos, A.M. et al. (1992) Sci
14 Baron, M. et al. (1992) Bioche
15 Patthy, L. (1987) J. Mol. Biol
16 Sharp, P.A. (1981) Cell 23, 64
17 Barton, G.J. and Sternberg, M
18 Devereux, J. et al. (1984) Nuc
19 Reid, K.B. and Day, A.J. (198
20 Barlow, P.N. et al. (1991) Bio
21 Idzerda, R.L. et al. (1990) J. E
22 Cosman, D. et al. (1990) Tre
23 Goodwin, R.G. et al. (1990) C
24 Bazan, J.F. (1990) Proc. Natl
25 Fukunaga, R. et al. (1991) EM
26 Tappin, M.J. et al. (1989) Eur
27 Cooke, R.M. et al. (1987) Na
28 Handford, P.A. et al. (1990) E

sequence of a placental cytoplasmic phospho-tyrosine phosphatase[101,102]. Subsequently PTPase activity has been shown for the membrane proximal cytoplasmic domain of CD45 but as yet not for the COOH-terminal domain[103]. Subsequently, other sequences have been identified with similarities to these sequences by cross-hybridization with cDNA probes and these include the LAR protein that is a cell surface protein with three IgSF domains and eight FN type IIISF domains on the extracellular side, although this protein is not expressed widely on leucocytes[104]. Sequences with many similarities to the PTPaseSF domains have been identified in *Drosophila*[105]. Some examples of the sequences are shown in Fig. 33. The second domain in CD45 is unusual in comparison with all other members of the PTPaseSF in that it contains an insertion of 19 amino acids with a very high content of acidic and Ser residues. The Ser residues may be phosphorylated by Ser kinases to produce an extremely negatively charged region of sequence.

The only complete genomic structure at present in the PTPaseSF is for mouse CD45[106]. This shows that the region illustrated in Fig. 33 is encoded by 6 or 8 exons for each domain. However, the ends of the domains as defined from the sequence similarities do not correspond to the ends of exons with the same phase of intron/exon boundaries. The genetic origin of these domains is unclear.

References

1 Dayhoff, M.O. et al. (1983) Methods Enzymol. 91, 524–45.
2 Williams, A.F. and Barclay, A.N. (1988) Annu. Rev. Immunol. 6, 381–405.
3 Baron, M. et al. (1991) Trends Biochem. Sci. 16, 13–17.
4 Pearson, W.R. and Lipman, D.J. (1988) Proc. Natl Acad. Sci. USA 85, 2444–2448.
5 Patthy, L. (1990) Cell 61, 13–14.
6 Williams, A.F. (1987) Immunol. Today 8, 298–303.
7 Williams, A.F. et al. (1989) Cold. Spring. Harb. Symp. Quant. Biol. LIV, 637–647.
8 Ryu, S.E. et al. (1990) Nature (London) 348, 419–426.
9 Wang, J. et al. (1990) Nature (London) 348, 411–418.
10 Driscoll, P.C. et al. (1991) Nature (London) 353, 762–765.
11 Holmgren, A. and Branden, C.I. (1989) Nature (London) 342, 248–251.
12 Amzel, L.M. and Poljak, R.J. (1979) Annu. Rev. Biochem. 48, 961–997.
13 de Vos, A.M. et al. (1992) Science 255, 306–312.
14 Baron, M. et al. (1992) Biochemistry 31, 2068–2073.
15 Patthy, L. (1987) J. Mol. Biol. 198, 567–577.
16 Sharp, P.A. (1981) Cell 23, 643–646.
17 Barton, G.J. and Sternberg, M.J. (1987) J. Mol. Biol. 198, 327–337.
18 Devereux, J. et al. (1984) Nucl. Acids Res. 12, 387–395.
19 Reid, K.B. and Day, A.J. (1989) Immunol. Today 10, 177–180.
20 Barlow, P.N. et al. (1991) Biochemistry 30, 997–1004.
21 Idzerda, R.L. et al. (1990) J. Exp. Med. 171, 861–873.
22 Cosman, D. et al. (1990) Trends Biochem. Sci. 15, 265–270.
23 Goodwin, R.G. et al. (1990) Cell 60, 941–951.
24 Bazan, J.F. (1990) Proc. Natl Acad. Sci. USA 87, 6934–6938.
25 Fukunaga, R. et al. (1991) EMBO J. 10, 2855–2865.
26 Tappin, M.J. et al. (1989) Eur. J. Biochem 179, 629–637.
27 Cooke, R.M. et al. (1987) Nature (London) 327, 339–341.
28 Handford, P.A. et al. (1990) EMBO J. 9, 475–480.

29 Rebay, I., et al. (1991) Cell 67, 687–699.
30 Constantine, K.L. et al. (1991) Biochemistry 30, 1663–1672.
31 Benian, G.M. et al. (1989) Nature (London) 342, 45–50.
32 Labeit, S. et al. (1990) Nature (London) 345, 273–276.
33 Suzuki, S. and Naitoh, Y. (1990) EMBO J. 9, 757–763.
34 Bjorkman, P.J. et al. (1987) Nature (London) 329, 506–512.
35 Leahy, D.J. et al. (1992) Cell 68, 1145–1162.
36 Piggott, R. and Power, C., (1992) The Adhesion Molecule FactsBook, Academic Press, London.
37 Erle, D.J. et al. (1991) J. Biol. Chem. 266, 11009–11016.
38 Takada, Y. and Hemler, M.E. (1989) J. Cell Biol. 109, 397–407.
39 Hemler, M.E. (1990) Annu. Rev. Immunol. 8, 365–400.
40 Drickamer, K. (1988) J. Biol. Chem. 263, 9557–9560.
41 Patthy, L. (1988) J. Mol. Biol. 202, 689–696.
42 Lasky, L.A. et al. (1989) Cell 56, 1045–1055.
43 Hoyle, G.W. and Hill, R.L. (1991) J. Biol. Chem. 266, 1850–1857.
44 Doege, K. et al. (1987) J. Biol. Chem. 262, 17757–1767.
45 Halberg, D.F. et al. (1988) J. Biol. Chem. 263, 9486–9490.
46 Collins, T. et al. (1991) J. Biol. Chem. 266, 2466–2473.
47 Johnston, G.I. et al. (1990) J. Biol. Chem. 265, 21381–21385.
48 Ezekowitz, R.A. et al. (1990) J. Exp. Med. 172, 1785–1794.
49 Weis, W.I. et al. (1992) Science 254, 1608–1615.
50 Frigeri, L.G. et al. (1990) J. Biol. Chem. 265, 20763–20769.
51 Hickey, M.J. et al. (1989) Proc. Natl Acad. Sci. USA 86, 6773–6777.
52 Schneider, R. and Schweiger, M. (1991) Oncogene 6, 1807–1811.
53 Wenger, R.H. et al. (1988) Biochem. Biophys. Res. Commun. 156, 389–395.
54 Braun, T. et al. (1991) EMBO J. 10, 1885–1890.
55 Suzuki, N. et al. (1990) Proc. Natl Acad. Sci. USA 87, 8711–8715.
56 Krantz, D.D. et al. (1991) J. Biol. Chem. 266, 16801–16807.
57 Perin, J.P. et al. (1987) J. Biol. Chem. 262, 13269–13272.
58 Aruffo, A. et al. (1990) Cell 61, 1303–1313.
59 Miyake, K. et al. (1990) J. Exp. Med. 172, 69–75.
60 Yamamato, T. et al. (1984) Cell 39, 27–38.
61 Südhof, T.C. et al. (1985) Science 228, 815–822.
62 Esser, V. et al. (1988) J. Biol. Chem. 263, 13282–13290.
63 Shevach, E.M. and Korty, P.E. (1989) Immunol. Today 10, 195–200.
64 Williams, A.F. (1991) Cell Biol. Int. Reports 15, 769–777.
65 LeClair, K.P. et al. (1986) EMBO J. 5, 3227–3234.
66 Deckert, M. et al. (1992) J. Immunol. 148, 672–677.
67 Williams, A.F. et al. (1988) Immunogenetics 27, 265–272.
68 Orr, H.T. et al. (1979) Biochemistry 18, 5711–5720.
69 Brown, J.H. et al. (1988) Nature (London) 332, 845–850.
70 Simister, N.E. and Mostov, K.E. (1989) Nature (London) 337, 184–187.
71 Beck, S. and Barrell, B.G. (1988) Nature (London) 331, 269–272.
72 Bjorkman, P.J. and Parham, P. (1990) Annu. Rev. Biochem. 59, 253–288.
73 Lawlor, D.A. et al. (1990) Annu. Rev. Immunol. 8, 23–63.
74 Leung, J.O. et al. (1985) J. Biol. Chem. 260, 12523–12527.
75 Sehgal, A. et al. (1988) Mol. Cell. Biol. 8, 3160–3167.

76 Strosberg, A.D. (1991) Eur. J.
77 Dohlman, H.G. et al. (1991)
78 Freeman, M. et al. (1990) Proc
79 Freeman, M. et al. (1991) Proc
80 Huang, H.J. et al. (1987) Proc
81 Reth, M. (1989) Nature (Lond
82 Letourneur, F. and Klausner,
83 Irving, B.A. and Weiss, A. (19
84 Rebbe, N.F. et al. (1991) Proc
85 Grundmann, U. et al. (1990)
86 Andria, M.L. et al. (1991) J. In
87 Wright, M. et al. (1993) Man
88 Levy, S. et al. (1991) J. Biol.
89 Oren, R. et al. (1990) Mol. C
90 Hupp, K. et al. (1989) J. Imm
91 Rubien, J.K. et al. (1989) In
University Press, Oxford, pp
92 Rudd, C.E. et al. (1989) Imm
93 Samelson, L.E. et al. (1990)
94 Reth, M. et al. (1991) Immu
95 Burkhardt, A.L. et al. (1991)
96 Ullrich, A. and Schlessinger
97 Cantley, L.C. et al. (1991) C
98 Hanks, S.K. et al. (1988) Sci
99 Yarden, Y. and Ullrich, A. (
100 Thomas, M.L. et al. (1985)
101 Charbonneau, H. et al. (198
102 Tonks, N.K. et al. (1988) Bi
103 Streuli, M. et al. (1990) EM
104 Streuli, M. et al. (1988) J. E
105 Streuli, M. et al. (1989) Pro
106 Hall, L.R. et al. (1988) J. Im
107 George, D.G. et al. (1990)
108 Williams, A.F. and Barclay,
Alt, F.W. and Rabbitts, T.H
109 Holmes, W.E. et al. (1991)
110 Gaugitsch, H.W. et al. (199
111 Flajnik, M.F. et al. (1991) I
112 Rippmann F. et al. (1991)

76 Strosberg, A.D. (1991) Eur. J. Biochem. 196, 1–10.
77 Dohlman, H.G. et al. (1991) Annu. Rev. Biochem. 60, 653–688.
78 Freeman, M. et al. (1990) Proc. Natl Acad. Sci. USA 87, 8810–8814.
79 Freeman, M. et al. (1991) Proc. Natl Acad. Sci. USA 88, 4931–4935.
80 Huang, H.J. et al. (1987) Proc. Natl Acad. Sci. USA 84, 204–208.
81 Reth, M. (1989) Nature (London) 338, 383–384.
82 Letourneur, F. and Klausner, R.D. (1992) Science 255, 79–82.
83 Irving, B.A. and Weiss, A. (1991) Cell 64, 891–901.
84 Rebbe, N.F. et al. (1991) Proc. Natl Acad. Sci. USA 88, 5192–5196.
85 Grundmann, U. et al. (1990) DNA Cell Biol 9, 243–250.
86 Andria, M.L. et al. (1991) J. Immunol. 147, 1030–1036.
87 Wright, M. et al. (1993) Manuscript in preparation.
88 Levy, S. et al. (1991) J. Biol. Chem. 266, 14597–14602.
89 Oren, R. et al. (1990) Mol. Cell. Biol. 10, 4007–4015.
90 Hupp, K. et al. (1989) J. Immunol. 143, 3787–3791.
91 Rubien, J.K. et al. (1989) In Leukocyte Typing IV, (Knapp, W. et al., eds) Oxford University Press, Oxford, pp.51–54.
92 Rudd, C.E. et al. (1989) Immunol. Rev. 111, 225–266.
93 Samelson, L.E. et al. (1990) Proc. Natl Acad. Sci. USA 87, 4358–4362.
94 Reth, M. et al. (1991) Immunol. Today 12, 196–201.
95 Burkhardt, A.L. et al. (1991) Proc. Natl Acad. Sci. USA 88, 7410–7414.
96 Ullrich, A. and Schlessinger, J. (1990) Cell 61, 203–212.
97 Cantley, L.C. et al. (1991) Cell 64, 281–302.
98 Hanks, S.K. et al. (1988) Science 241, 42–52.
99 Yarden, Y. and Ullrich, A. (1988) Annu. Rev. Biochem. 57, 443–478.
100 Thomas, M.L. et al. (1985) Cell 41, 83–93.
101 Charbonneau, H. et al. (1988) Proc. Natl Acad. Sci. USA 85, 7182–7186.
102 Tonks, N.K. et al. (1988) Biochemistry 27, 8695–8701.
103 Streuli, M. et al. (1990) EMBO J. 9, 2399–2407.
104 Streuli, M. et al. (1988) J. Exp. Med. 168, 1523–1530.
105 Streuli, M. et al. (1989) Proc. Natl Acad. Sci. USA 86, 8698–8702.
106 Hall, L.R. et al. (1988) J. Immunol. 141, 2781–2787.
107 George, D.G. et al. (1990) Methods Enzymol. 183, 333–351.
108 Williams, A.F. and Barclay, A.N. (1989) In Immunoglobulin Genes, (Honjo, T. Alt, F.W. and Rabbitts, T.H. eds) Academic Press, London, pp.361–387.
109 Holmes, W.E. et al. (1991) Science 253, 1278–1280.
110 Gaugitsch, H.W. et al. (1991) Eur. J. Immunol. 21, 377–383.
111 Flajnik, M.F. et al. (1991) Immunogenetics 33, 295–300.
112 Rippmann F. et al. (1991) EMBO J 10, 1053–1059.